ENABLING PRECISION MEDICINE BY INTEGRATING MULTI-MODAL BIOMEDICAL DATA

A Dissertation
Presented to
The Academic Faculty

By

Li Tong

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
Department of Biomedical Engineering

Georgia Institute of Technology and Emory University

December 2020

ENABLING PRECISION MEDICINE BY INTEGRATING MULTI-MODAL BIOMEDICAL DATA

Approved by:

Dr. May D. Wang, Advisor
Department of Biomedical Engineering
Georgia Institute of Technology and Emory University

Dr. Wei Sun
Department of Biomedical Engineering
Georgia Institute of Technology and Emory University

Dr. Omer T. Inan
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Nikhil K Chanani
School of Medicine
Emory University

Dr. Shriprasad Deshpande
School of Medicine
Children's National Health System and George Washington University

Date Approved: November 23, 2020

We are all in the gutter, but some of us are looking at the stars

Oscar Wilde

Dedicated to my parents

For all their love, support, and encouragement

ACKNOWLEDGEMENTS

There are no proper words to convey my deep gratitude and respect for my thesis and research advisor, Dr. May D. Wang. She has inspired me to become an independent researcher and guided me through the challenges and hard times. She also demonstrated what a brilliant and hard-working scientist could accomplish. My sincere thanks must also go to the members of my thesis advisory and exam committee: Dr. Wei Sun, Dr. Omer T Inan, Dr. Nikhil Chanani, and Dr. Shriprasad Deshpande. They generously gave their time to offer me valuable comments toward improving my work. In particular, Dr. Deshpande has not only provided research data as a collaborator but also provided me constructive suggestions and guidance, which helped me sharpened the insights of my thesis.

I am most grateful to the collaborators for providing me the research data and their expertise to my scientific and technical problems: Dr. Kevin E. Woods, Dr. Deshpande, and Dr. Weida Tong. There is no way to express how much it meant to me to have been a member of BioMIBLab. These brilliant friends and colleagues inspired me over the many years: Hang Wu, Ryan Hoffman, Yuanda Zhu, Ying Sha, Wenqi Shi, Anirudh Choudhary, Felipe Giuste, Hamid Hassanzadeh, Dr. Po-Yen Wu, Dr. Janani Venugopalan, and Dr. John Phan, and all the other current and former BioMIB Lab grad students and visitors that I know.

I cannot forget friends who went through hard times together, cheered me on, and celebrated each accomplishment: Hang Wu, Ying Sha, Shiying Wei, Yao Chen, Nan Xie, Derrick Chu, Xiao Zang, and Yilun Han.

Finally, I want to deeply thank my parents, Shige Tong and Qinghong Zhou, for their unconditional trust, timely encouragement, and endless patience. It was their love that raised me up again when I got weary and helped me get through this challenging period of time in the most positive way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

With the advancement of biotechnologies such as high-throughput sequencing, a massive amount of multi-modal biomedical data has been generated at an unprecedented speed and volume every year. However, extracting information and obtaining knowledge from these multi-modal biomedical data remains a major challenge in research and clinical applications. Multiple computational approaches were proposed for multi-modal biomedical data integration, aiming to combine data from disparate sources to increase the value of data and improve data integrity. Multi-modal biomedical data were hypothesized to contain both dependent and independent information based on multi-view learning's consensus and complementary principles. For modalities with independent information or few connections (e.g., genetic factors vs. environmental factors), the complementary principle was utilized to integrate data from different modalities by concatenating the hidden features learned with independent feature representation. Thus, the unique information in each data modality can jointly contribute to the final decision. The proposed framework has been applied to integrate electronic health records (EHRs) with MRI Imaging and single nucleotide polymorphisms (SNPs) data for improved prediction of Alzheimer's Disease. For modalities with dependent information (e.g., multi-omics data), the complex interactions between modalities were modeled implicitly with the consensus principle. As features from dependent modalities are connected by either association or causal relationships, they can be integrated by the consensus principle to improve the robustness and eliminate inconsistencies. A consensus regularization was achieved by requiring the features encoded from various modalities of the same subject to consent in a common feature space. The proposed frameworks have been applied to integrate multi-omics data (e.g., mRNA expression, DNA methylation, miRNA expression, and copy number variations) for improved breast cancer overall survival prediction. Generalized data integration models such as autoencoder-based semi-supervised learning frameworks have also been explored to improve computer-aided

decision support performance. By integrating multi-modal biomedical data with the proposed frameworks, the healthcare quality is expected to be improved with a more comprehensive evaluation of the patient.

CHAPTER 1
INTRODUCTION

Delivering predictive, precise, participatory, preventive, and personalized health, abbreviated as p-Health, is the primary goal of future healthcare systems that can significantly improve care quality while reducing cost. To accomplish this goal, researchers are developing translational data analytics pipelines to jointly assess and validate the in-clinic Electronic Health Records (EHRs), out-of-clinic Personal Health Records (PHRs) [1], and high-throughput genomic data. A typical pipeline consists of six steps: data collection, data quality control, feature extraction, knowledge modeling, decision making, and action-taking. However, how to integrate multi-modal biomedical data in an efficient way remains a major challenge. This chapter will review the challenges and opportunities in data integration analytics at three levels (i.e., raw data, feature, and decision levels) from clinical informatics to genomics for p-Health as shown in Figure 1.1.

Clinical informatics was first introduced in the 1950s, when the US National Bureau of Standards, the US Air Force used digital computers to develop expert systems (such as MYCIN and Internist-I [2]) and computerized medical records management system. In 1959, Ledley et al. [3] published in Science that discussed the idea of using computational reasoning to aid medical diagnostic processes for the first time. EHR systems improve the communication among physicians, providers, and patients, the quality of clinical decisions, and the delivery of cares significantly in hospital routine practices [4]. However, due to the low EHR adoption rate, clinical informatics progress was slow. In 2008, the American Hospital Association (AHA) Annual Survey reported that only 13.4% of the non-Federal acute care hospitals in the US had adopted basic or comprehensive EHR systems and only 1.6% have an EHR system with clinical decision support [5]. With the new policy "the Health Information Technology for Economic and Clinical Health (HITECH) Act 2009", in 2015,

the EHR system adoption rate has increased to about 80% in the US, and 34.4% of these hospitals have EHR systems with clinical decision support [5]. Recently, the advancement of low-cost sensors embedded in mobile phones or wearable devices has enabled out-of-clinic care captured in PHRs that expand EHRs. The continuous monitoring of a person's daily physiology and communication enables personalized preventive and predictive health by disease early warnings and participatory patient education that bridges the gap between in-clinic and out-of-clinic care for p-Health. We loosely use EHRs to cover both traditional EHR and PHR for clinical informatics within the context of this dissertation.

After Human Genome Project was finished in 2001 [6], using genomics to customize clinical care becomes possible. Genomics focuses on discovering the structure and function of genomes, which is the complete set of DNA in an organism. Based on the National Human Genome Research Institute (NHGRI) definition, "Genomic Medicine" is an emerging medical discipline that uses genomic information of an individual for his/her clinical care and health outcomes. The NHGRI "Genomics" covers the study of direct information about DNA or RNA, excluding the study of downstream derived products (e.g., proteomics, metabolomics). DNA sequences can capture genomic variations at single nucleotide level like single nucleotide polymorphisms (SNPs) [7] or chromosome level like structure variations (SVs) [8]. RNA sequences contain genomic variations as gene expressions or alternative splicing events [9]. Because human diseases are complex interactions between genotypes and environment [10], incorporating molecular level information such as genetic variations is essential for precision medicine to speed up the accomplishment of p-Health. The unique genetic information of individuals can reveal the disease status and responses to treatments (e.g., more than 80% of rare diseases are related to genetic mutations, and genomics can play an important role in the diagnosis [11]). The genomic variations detected in DNA or RNA sequences are genotypes, a complement to disease phenotypes in genomic medicine. With the development of high-throughput next-generation sequencing (NGS) technologies, genomic data such as the whole genome of an individual

can be sequenced for as little as $1,000 in a few days [12].

Integrating "clinical informatics" and "genomic medicine" presents us with challenges including data harmonization, data quality control, and advanced data analytics to build an integrated intelligent clinical decision support system for p-Health (Figure 1.1). To solve these challenges, we can perform data integration at three different levels of the six-step data-analytic pipeline: raw data level, feature level, and decision level (Figure 1.1).

1.1 Clinical Informatics

Clinical informatics uses data analytics to gain new insight from individual and population health records and to improve clinical decision making by combining data-derived knowledge and domain expert knowledge (Figure 1.2). Typical data sources include conventional personal health records and public health data.

Conventional individual health records contain various data types, from structured billing codes to unstructured clinical notes, which is the first challenge in clinical informatics. Throughout the hospital stay of patients, administrative information (e.g., demographics, gender information, and diagnostic codes, for billing and public health reporting purposes), auxiliary clinical data (e.g., lab tests, medical and radiological imaging, medication, genetics, continuous physiological data such as bedside monitoring data in intensive care units), and unstructured clinical notes record comprehensive information about patients' clinical status and outpatient care information. The public health data set is more abstract and often contains aggregated statistics of diseases in geographical regions and times. For example, the national center for health statistics (NCHS) collects 2.6 million death certificates each year that record demographics, causes of death of the US population. Recently, billions of mobile phones and wearable sensors (e.g., FitBit) and behavior imaging enable routine health monitoring in outpatient care. PHRs collect data from these applications and provide a more personalized evaluation of patients. Within the context of this dissertation, we use EHRs to include EHRs, physiological and imaging data, and loosely group PHRs together

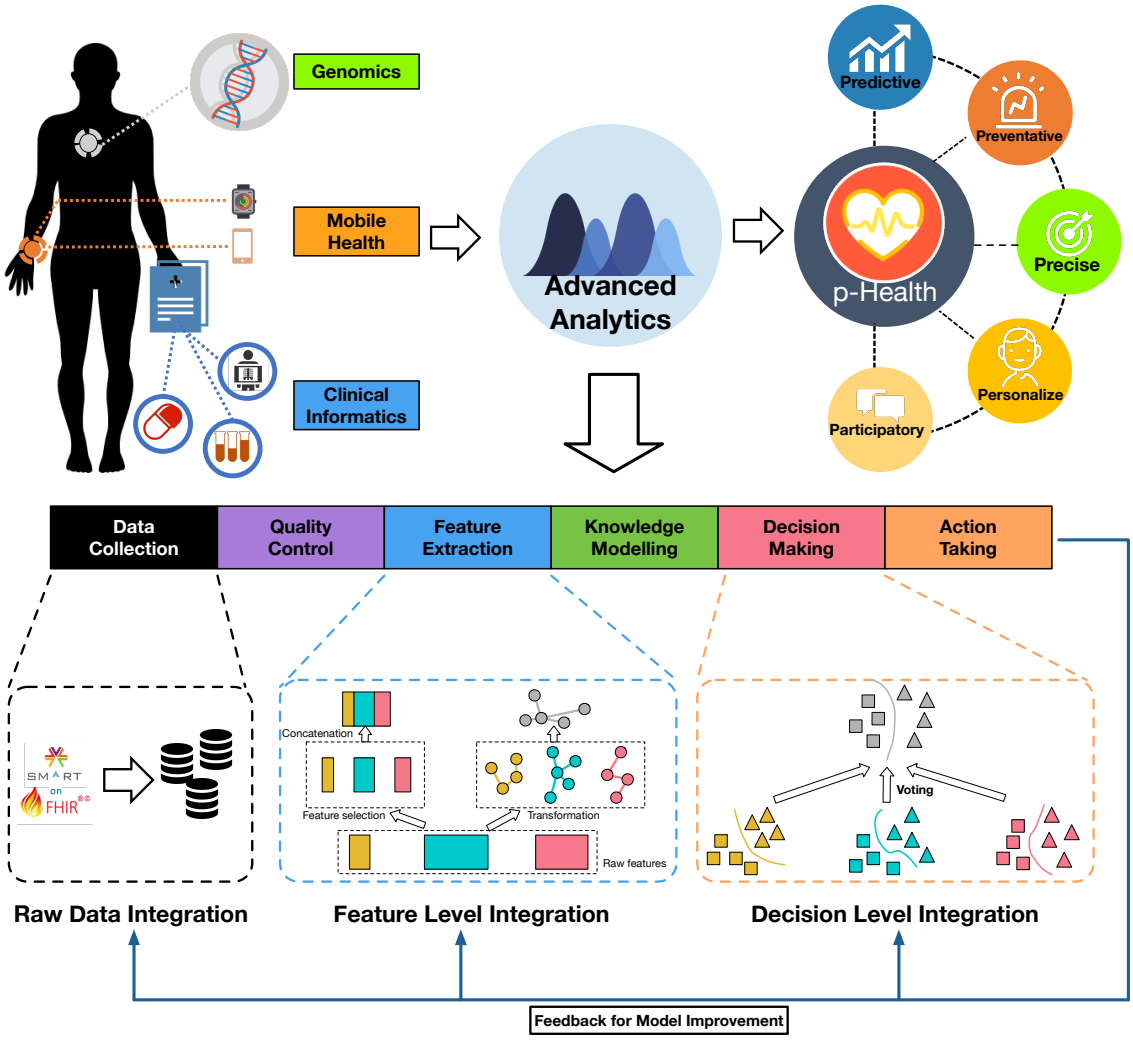


Figure 1.1: Integrating clinical informatics with genomics and mobile health enables p-Health through advanced data analytics. The typical data analysis pipeline consists of data collection, quality control, feature extraction, knowledge modeling, decision making, and action-taking. Data integration can be achieved along this pipeline at the raw data level, feature level, and decision level, respectively.

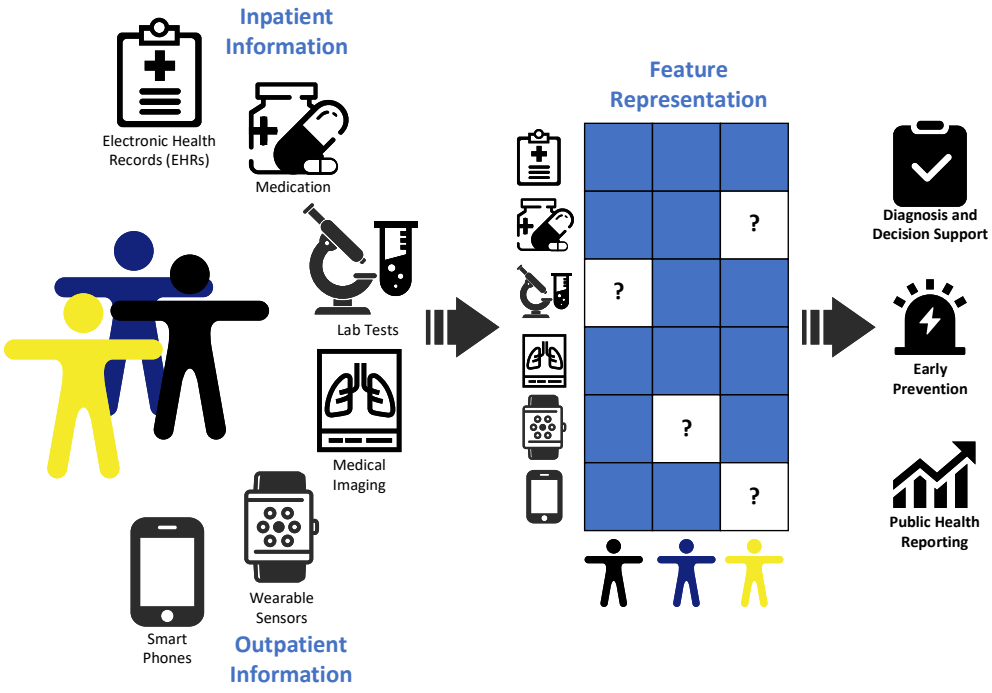


Figure 1.2: A overview of clinical informatics. In-clinic and out-of-clinic health information contribute to the primary data sources for clinical informatics. Data from different sources are combined to form the feature representation of each patient, for later clinical endpoints such as clinical outcome prediction, early prevention, and public health reporting.

with EHRs to become "extended" EHRs.

Other data sources that attribute to health include ontology, such as Unified Medical Language System (UMLS), SNOMED CT [13]. Physicians and clinical researchers use nodes to represent medical concepts and edges between nodes to encode the relationships between concepts such as what diseases can lead to specific symptoms, what medications can help alleviate certain symptoms.

Analyzing these data has benefits to both population and individual care for p-Health. At the population care level, health policymakers are interested in understanding epidemics from a statistical point of view across spatial and temporal dimensions [14]. Aggregated EHRs help explore the evolution and spreading of diseases for resource allocation, which may be used to predict and prevent the outbreak of epidemics. Death certificates may contain the causes of deaths for each death case and help understand the diseases spread in the nation. At the individual care level, diagnosing patients' conditions and deciding prognosis are two crucial steps. Clinical informatics can provide decision support for in-clinic physicians and care providers [15], where clinical outcome (e.g., hospital readmission and mortality) prediction helps prevent adverse in-clinic clinical events. Mobile sensor data prediction gives early warnings of medical conditions and provides physicians with out-of-clinic participatory health. Besides prediction, grouping similar patients' record [16] and simulation [17] can suggest reference treatment options to assist in the clinical action-taking.

Due to privacy and other health regulatory issues, only a limited number of datasets are publicly available. Comprehensive data sets may be obtained through research collaboration with specific clinical institutions. As the data available are only multiple types of observations, to truly achieve personalized and precise health [17], it is critical to expanding to genome medicine[18], which is presented in the next section. However, clinical informatics still falls short in the challenge of precision and personalized medicine.

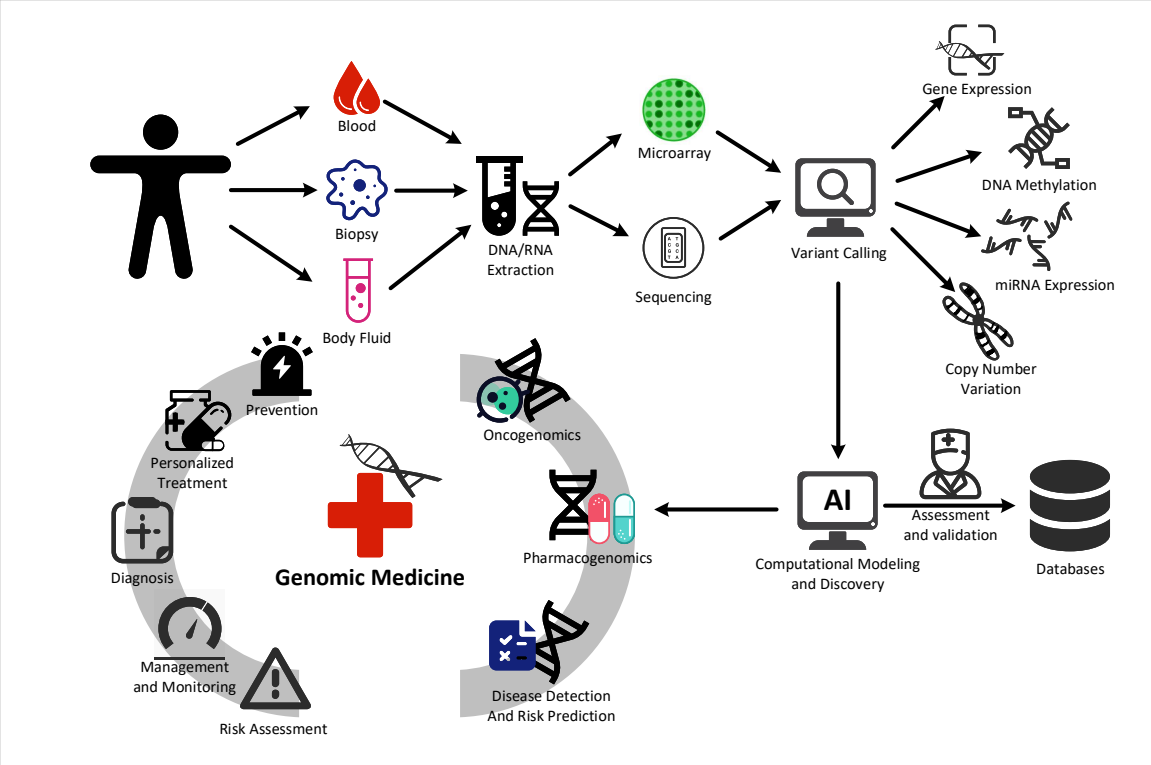


Figure 1.3: An overview of genomic medicine. DNAs/RNAs are firstly extracted from the patient's blood, biopsy, or body fluid samples, and then genotyped by either microarray or sequencing. The raw genomic data are further analyzed with bioinformatics approaches to generate a list of genomic variants. AI-based computational modeling is applied for disease-related genomic variants discovery with the guidance of current knowledge bases. The identified genomic biomarkers will be evaluated and validated by domain experts and enrich the knowledge bases. Oncogenomics, pharmacogenomics, and disease risk prediction are three major components of genomic medicine, enabling disease prevention, diagnosis, personalized treatment, risk assessment, and health management and monitoring.

1.2 Genomic Medicine

Genomic medicine aims at utilizing an individual's genomic information for preventing, diagnosing, and treating diseases in clinical care and p-Health (Figure 1.3). Human diseases are results of complex interactions of genotypes and environmental factors [10]. Personalized information embedded in genomics can help physicians predict disease risk factors and patients' responses to treatments. For example, because specific inherited mutations in tumor suppressor genes BRCA1 and BRCA2 increase the risk of female breast and ovarian cancers [19], genetic tests of BRCA1 and BRCA2 mutations can prevent, early-diagnose,

and treat breast and ovarian cancers. Genomic medicine has been mainly applied to three areas: oncogenomics, pharmacogenomics, and disease risk prediction. Oncogenomics and pharmacogenomics are the two most studied applications of personalized genomics [20]. Disease risk prediction is especially beneficial for those diseases with genetic markers of high clinical impacts.

Oncogenomics is a study for cancer–related genes. As a well-recognized genetic disease [21], various cancers have been studied extensively with genomics to identify cancer-related genetic variations for personalized diagnosis and treatment [22]. For example, Nguyen et al. have applied personal genomics to relate the outcome after breast-conserving therapy with breast cancer subtypes [23]. Zheng et al. have associated genetic variations with prostate cancer [24]. Pharmacogenomics is the study of how a person's genome affects his/her response to drugs (i.e., the relationship between genetic variations and drug responses). It can guide the physicians to choose the optimal medicine for patients based on their genetic variations to minimize adverse drug reactions and to maximize effectiveness [20] for personalized care. For example, the chemotherapeutic agent Vemurafenib targets melanoma with BRAF V600E mutation, where the efficacy and adverse events have been evaluated [25]. Disease risks vary with the patient's genetic variations. For example, 80% of rare diseases have genetic origins, which makes personalized genomics essential for early screening, diagnosis, and individualized treatment for rare diseases [11]. Multiple rare diseases have genetic biomarkers with sufficiently high clinical impact, and more are discovered each year. Combining genetic biomarkers using genome-wide association studies (GWAS) with environmental factors, the risk factors of some diseases can be thoroughly assessed for earlier intervention and better treatment. Various databases and projects have been established for genetic variants discovery for general purposes, specific diseases, and specific populations. The first group of projects aims to discover genetic variants and to understand the functions. To identify the common genetic variations and to discover novel ones in the human genome, 1000 Genomes Project [26], followed by

10,000 human genomes project [27], sequenced a large number of human genomes. These projects provide a base for the clinical use of genetic variations. The encyclopedia of DNA elements project (ENCODE) [28] aims to discover the functional elements in the human genome to make sense of genomic data. UK Biobank [29] connects genetic variants with a wide range of diseases and outcomes by providing EHR (with imaging) and genomic data, which is a perfect resource for integrated data analytics.

The second group of projects aims to understand and treat specific diseases. These projects include the cancer genome atlas project (TCGA) [30], Alzheimer's Disease neuroimaging initiative (ADNI) [31], and Parkinson's progression markers initiative (PPMI) [32]. TCGA project focuses on utilizing cancer genomics to improve the understanding, prevention, diagnosis, and treatment of various cancers. Similarly, ADNI and PPMI projects are developed for two of the most prevalent neurodegenerative diseases, Alzheimer's disease and Parkinson's disease, respectively. These disease-oriented databases collect multi-omics data, clinical information (EHR), and medical imaging data (magnetic resonance imaging (MRI), positron emission tomography (PET), and pathological images) for specific diseases. The multi-modal data collected by these databases enable data integration for a better understanding of the disease.

The third group of projects aims to stratify medicine for specific populations. One example is the minority health genomic and translational research bio-repository database (MH-GRID) [33]. Motivated by the fact that high blood pressure affects African Americans more than other racial groups, MH-GRID contains data collected from over a thousand African Americans across the US. Besides genetic data, MH-GRID also collects health-related information such as diet, sleep, body mass index, stress, access to healthy food and parks. Besides publicly available databases and projects, various bioinformatics tools and pipelines have also been developed for DNA and RNA analysis in personalized genomic medicine. DNAs and RNAs can be first sequenced by next-generation sequencing (NGS) platforms and then analyzed by bioinformatics approaches. Developed from the Sager

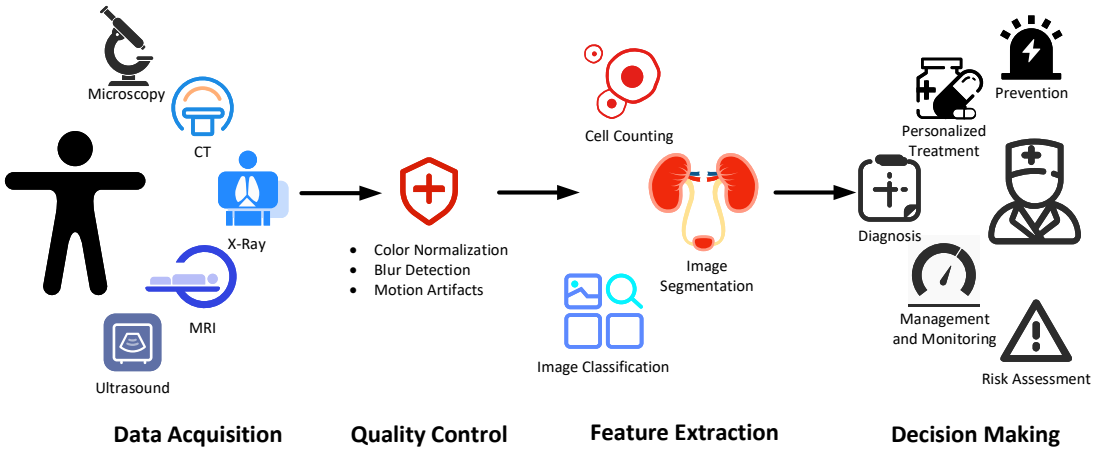


Figure 1.4: An overview of medical imaging informatics. Medical imaging informatics aims to enable quantitative and robust analysis of medical images by computational methods.

method's chain-termination, NGS techniques parallelize the sequencing process and produce millions of sequences concurrently [34] by using sequencing-by-synthesis, and have significantly lowered the cost of DNA sequencing compared to the Sanger method [34]. The first step of genomic data analysis is sequencing data quality control to remove potential artifacts and low-quality reads. The sequenced short reads passed quality control are then aligned to the reference genome. The aligned reads are further analyzed to identify various genetic variations like SNPs and copy number variations (CNVs). Based on the NHGRI genomic definition, genetic variations at the transcription level can be characterized by RNA-seq to obtain gene expression, alternative splicing, and gene fusion. That is, RNA sequencing (RNA-seq) reads are first aligned to the genome (spliced alignment) or transcriptome (un-spliced alignment) and are then quantified and normalized to get gene expression levels. RNA-seq can capture major transcription-level regulation events like alternative splicing [9]. These innovations in sequencing technologies and bioinformatics pipelines pave the way for personalized genomic medicine.

1.3 Medical Imaging Informatics

Medical imaging informatics aims to improve imaging-based diagnosis's efficiency and accuracy through digital imaging processing and machine learning (Figure 1.4). A typical computer-aided diagnosis (CAD) system for medical imaging analysis consists of four major components: 1) image quality control, 2) feature extraction at the pixel, object, and semantic levels, 3) predictive modeling using imaging features, and 4) model visualization for interactive discovery [35].

Image quality control is an essential step for medical imaging analysis. Typical quality issues for medical images include batch effects and artifacts. For example, the batch effects for pathological images are caused by different clinical sites with various platforms and slide preparation protocols. Besides batch effects, pathological images also suffer from artifacts, including tissue folds, blurred regions, pen marks, and shadows. The batch effects and image artifacts have unpredictable effects on image segmentation, classification, and other quantitative image analysis tasks. Researchers have developed multiple techniques, including color normalization, scale normalization, and blur detection, to eliminate or correct the batch-effect and image artifacts of WSIs.

Feature extraction is another essential step to analyze medical images quantitatively. Conventional digital imaging processing techniques extract features from pathological images at pixel and object levels to capture the morphological properties [36]. Pixel-level feature extraction identifies the properties of color and texture for all image pixels. Color features are typically expressed with the color spread, prominence, and co-occurrence using statistics and frequencies of color histograms in different color spaces. Texture features quantify image sharpness, contrast, changes in intensity, and discontinuities or edges by measuring properties from gray-level intensity profiles. Object-level feature extraction requires the segmentation of cellular structures and captures the shape, texture, and spatial distribution of cellular structures in a WSI. Besides the features extracted from WSI, re-

searchers also proposed to integrate the pathological features with clinical features and genomic features for improved diagnosis [37].

However, the conventional feature extraction relies heavily on hand-crafted features, which limits the generalizability of the features. With the development of deep learning, the human-designed feature extraction has been replaced by feature representation with deep neural networks. For imaging data, the most popular feature representation network is the convolutional neural network (ConvNets). Deep ConvNets can learn efficient feature representation from a large amount of training data. By combining ConvNets with fully connected (FC) layers and a softmax layer for classification, the deep networks can be trained end-to-end and thus can learn both feature representation and classification from the training data. With the success of deep learning in natural images, deep neural networks like ConvNets have been applied to medical images including MRI (brain tumor [38]), CT (lung nodule [39]), and WSI (breast cancer [40], lung cancer [41], glioma [42], heart rejection [43], etc.).

With multi-modal biomedical data discussed above, how to combine various data modalities of the same patient becomes a practical yet challenging task. Valid data integration should provide a more comprehensive evaluation of the patient and enable personalized diagnosis and treatment. In the following chapters, we will discuss the practical challenges for data integration and present the proposed data integration methods.

CHAPTER 2
INTEGRATING MULTI-MODAL BIOMEDICAL DATA WITH FEATURE CONCATENATION

2.1 State-of-the-art biomedical data integration

Data integration aims to use multiple sources of information to better understand a system (Figure 1.2). Wong reviewed requirements and solutions to the data integration and warehouse at the raw data access level in biomedicine in EnsEMBL, GenoMax, and SRS [44]. Goble et al. introduced "a loose federation of bio-nations" to handle biomedical data sources' heterogeneity and the ontology in data integration [45]. More recently, Gomez-Cabrero et al. described how to deal with the diversity of existing omics data types and formats of large datasets to obtain new insights [46]. They have applied dynamic Bayesian networks, self-organizing maps, and network inference methods over popular data sources such as the 1000 Genomes Project, ENCODE, TCGA, and ImmGen. Ritchie et al. reviewed the -omic data integration by comparing meta-dimensional and multi-stage analysis [47]. Based on the current works for biomedical data integration, we identify the following four challenges for integrating EHRs and genomic data:

The first challenge is data collection and harmonization. New technologies, such as mobile sensors and DNA sequencing, all require modality-specific techniques to be stored, visualized, and analyzed by computing devices. Genomic technologies are emerging every year, ranging from microarray to third-generation sequencing, making genomic data heterogeneous with various formats and standards. Then NGS have different platforms such as Illumina, 454 sequencing, or SOLiD sequencing, leading to more variations. The development of third-generation sequencing by Pacific Biosciences [48] with much longer read length will add further variants to genomic data. Validating the data generated by these var-

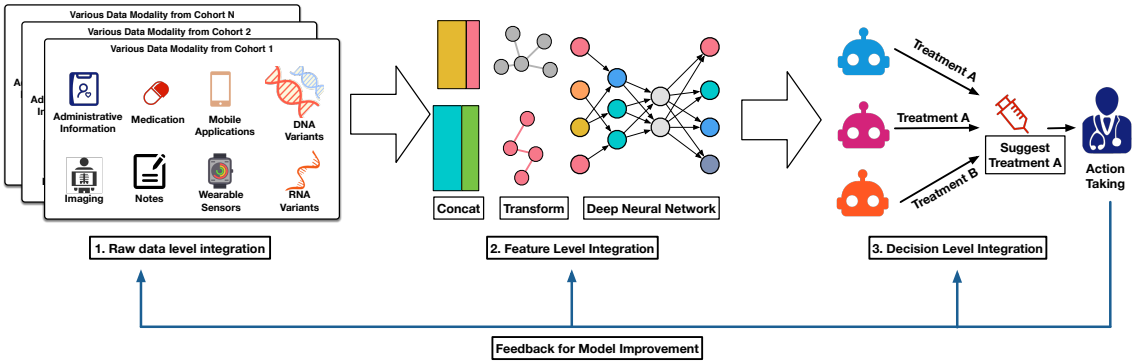


Figure 2.1: Summary of advanced data integration analytics. The multi-modal health data can be integrated at the raw data level, feature level, and decision level. At the raw data level, health data from different modalities and cohorts are harmonized. Features from different data can be integrated by concatenation, transformation, or deep learning methods at the feature level. At the decision level, the outputs of multiple models are combined for decision making and further action taking by the doctors.

ious technologies is challenging. How to harmonize different types of data and databases is also challenging. In clinical informatics, there are 186 commercial EHR system vendors in the U.S. Market, with Epic, Cerner, and MEDITECH being the top three with the highest market shares. Each of them has its specific data standard and format. As a result, multiple clinical institutions with different EHR vendors cannot exchange information directly [49]. The international standard body, Health Level Seven International (HL7) [50], has put a collaborative effort to develop a new EHR standard FHIR (Fast Healthcare Interoperability Resources), and its extensions SMART-on-FHIR to have data from different EHR interoperable. In genomic medicine, currently, FHIR enables the incorporation of genomic variations into EHRs. One example of establishing standards and APIs for genomic data storing, sharing, and clinical applications is SMART on FHIR Genomics [51]. Thus, FHIR app development presents challenges and opportunities.

The second challenge is data quality. In clinical informatics, EHR data can be unstructured and noisy with issues such as high percentages of missing values, errors, invalid data, and outliers. Thus, data quality control processes, such as missing data imputation, data conflict resolution, and data transformation, are needed. In genomic medicine, the DNA

sequencing quality is affected by sample contaminations and sequencing errors. Thus, numerous new bioinformatics approaches (e.g., FastQC) are developed to improve the sequencing quality before downstream analysis. In addition, to establish unified standards for genomic data processing for reliable clinical applications, the US Food and Drug Administration (FDA) coordinates the microarray quality control (MAQC) project [52] and the sequencing quality control (SEQC) project [53] for a comprehensive assessment of microarray and RNA-seq respectively.

The third challenge is to develop advanced data analytics to extract knowledge from EHR and genomic data. In clinical informatics, feature engineering constructs a meaningful set of representations for patients, including medications, lab tests, and procedures. The conventional analytics rely on hand-crafted features or input from domain experts, which are very limited in representation power. The new opportunities are on deep learning to find representations from extensive data. In genomic medicine, the "curse of dimensionality" (i.e., the feature dimension is significantly larger than the patient sample size) is a well-known challenge. For example, genome sequencing generates millions of genomic variations such as SNPs, CNVs, gene expression levels, and alternative splicing events for each sample. The sample size of the study may be several hundred. Directly applying traditional data analysis will result in ill-conditioned feature matrixes. One solution is to filter out irrelevant variants using feature selection in either a supervised or unsupervised fashion. Lack of interpretability is another major challenge for EHR and genomic data analytics. Predictive models built on EHR data have shown potential in predicting the hospital length of stay and readmission probability. However, deploying them in daily hospital practice is challenging because physicians cannot fully trust the model prediction. Most models only provide a final prediction without a clear explanation of patients' conditions lead to the prediction. Making sense of genomic data is also a bottleneck for translating genomic discoveries into clinical practice [54]. Current data-driven approaches are mostly based on mathematical models, statistical tests, and computational methods. Lack of biological and

clinical interpretation weakens the clinical impact of the novel genetic biomarkers discovered. Experimental validation is essential for translating discoveries from data mining into biological knowledge and further clinical care applications.

The fourth challenge concerns how to transfer knowledge to actionable decision making. Artificial intelligence (AI) started in the 1950s and has undergone the paradigm shift from symbolic logical reasoning to machine learning [55], including the recent success of deep learning [56] in automatic representation learning. Deep learning has the potential to provide highly accurate predictive modeling, and deep reinforcement learning shows significant breakthroughs in playing video games, dialogue systems, and other tasks. However, a major challenge for computation decision making in clinical informatics is that it is unethical to test different options and practically almost impossible to test two treatment options in one individual. We may use historical data to simulate the effect of each action applied to an individual, termed as counterfactual inference in the literature. However, to ultimately address the problem, we need to understand the causal relationship between medical events, treatment options, and diseases, for causality beyond correlation studies.

With these challenges presented, numerous works have been performed on multi-modal biomedical data integration. This section reviews state-of-the-art data integration analytics at the raw data level, feature level, and decision level to enable p-Health.

2.1.1 Raw Data Level Integration

Raw-data level integration happens in the data collection and storage stage, where different modalities of data are collected from different institutions and times for the same disease. The multi-modal data from more patients can enable machine learning algorithms to achieve better predictive models [57]. However, several challenges are also associated with raw data integration.

The main challenge is that different institutes may use different EHR vendors with different formats. Existing standards such as Health Level Seven International (HL7) and

Clinical Document Architecture (CDA) [50] cannot support the interexchange of data from different EHRs automatically for decision support systems. In contrast, manual data conversion case-by-case is both time-consuming and prone to errors. Besides, ensuring privacy and security in the raw data integration is also crucial [58]. To harmonize the data, HL7's has been developing Fast Healthcare Interoperability Resources (FHIR). FHIR is a new emerging standard that uses a resource-centric approach (as opposed to document-centric). FHIR specifies data elements into standardized healthcare data models and a set of application programming interfaces (APIs) for interacting and modifying these data models [59]. Thus, data resources can exist as online services rather than static files so that applications can read and write these resources in real-time. SMART-on-FHIR apps are based on FHIR but work outside of the constraints by many stakeholders' existing technical and security infrastructures. Thus, raw data level integration opportunities include extending FHIR Resources, implementing RESTful APIs, and SMART-on-FHIR apps to facilitate the raw data integration process. For example, the applications include death reporting, health record vendor translator, and mobile health apps [60]. Another example is the SMART on FHIR Genomics for establishing standards and APIs for genomic data storing, sharing, and clinical applications [51].

2.1.2 Feature Level Integration

The next step in the data analytics pipeline following data collection and quality control is feature extraction, and feature level integration includes both concatenation-based and model-based methods. Directly concatenating raw features to integrate different data modalities can introduce thousands of features. However, the challenges for integrating EHRs and genomics features at the raw feature-level are that the information represented by feature vectors may have varying representation power and noise. The computational challenges introduced by adding features together may lead to model under-fitting or non-convergence. Thus feature selection, such as L1 norm feature selection [61] and minimum-

redundancy maximum-relevancy (mRMR) feature selection [62], is needed.

Model-based feature integration often adopts an encoder-decoder framework to combine features from different modalities. For each modality, modality-specific encoders can map features of every data modality to a joint feature embedding space first. Features combined in the joint space will go through a single decoder for final reporting. Various machine learning algorithms have been researched to accomplish model-based integration. Multiple kernel learning (MKL) [63] learns a kernel transformation for each modality and then combines all kernel transformations with a weighted linear average. Canonical correlation analysis (CCA) [64] finds a new linear feature space, in which features from all modalities has maximum correlation when projected onto the new space. Probabilistic graphical model (PGM) [65] treats each feature as a random variable and perform statistical inference to obtain an integrated feature, which is also represented in the form of latent variables. These techniques are effective in several applications. However, the disappearance of individual modality's feature learning and feature integration may result in a meaningless feature representation.

An emerging advanced AI method in feature-level integration is deep learning that can combine feature extraction and prediction, learning a meaningful representation for high accuracy in the given task [56]. Multi-modal deep learning [66] is an early work using the restricted Boltzmann machine, while CCA is also extended with deep feedforward network in [67] for feature-level integration. In clinical research, the multi-modal analysis using deep learning is shown to improve model accuracy in medical imaging analysis [68].

The opportunities in advanced AI-based feature integration include constructing differentiable encoders and the interpretation of integrated features and predictive models, especially deep learning models. Differentiable encoders can extract features from multi-modal data (e.g., EHRs and genomic data) and train the whole pipeline with back propagation. On feature interpretation, despite the progress in building highly accurate predictive models, physicians will not trust black-box algorithms if they cannot understand what the features

contribute to the final prediction and how they can link features to the original physiological meaning. Perturbation analysis [69], influence functions [70], and visualization of neural network outputs are some pioneering work for model interpretation. Identifying important raw features that contribute to the final prediction or similar patients from historical data are two directions that may help us understand the behavior of current models.

2.1.3 Decision Level Integration

Decision level integration first generates multiple base models using each data modality independently as the training sets and then generates a final model by combining individual models trained in the first step.

The challenges of decision level integration are the construction of accurate base decision models and the combination of base models that can prevent model overfitting while adding more parameters.

The conventional decision-level combination uses simple majority, the weighted majority [71] (to predict protein fold recognition [72]), and ensemble learning. Ensemble learning designs special weighting methods to addresses the overfitting by adjusting base models with the resampled training set (bagging [73]), sequentially increasing weights of misclassified training data (boosting [74]). For example, the random forest classifier makes decisions based on the multiple decision trees it constructed from the resampled training set; graphical-model-based approaches such as Bayesian networks combines multi-omics data to better understand glioblastoma and breast cancer [75, 76]

In advanced AI methods for decision level integration, there exist three opportunities: construction of base models, increasing interaction between different data modalities, and reducing the gap between the decision and clinical action. First, designing accurate base models is critical for final predictive accuracy. For example, in the majority voting scheme, we need to choose one or more base models for each data modality from different models, including decision tree, k nearest neighbors (kNN), support vector machine, logistic regres-

sion, and neural network. For each base model, there are also multiple parameters to be tuned. The selection of base models and parameter tuning can be tedious and also prone to overfitting, especially for small sample sizes, so inventing efficient hyper-parameter tuning algorithms such as Bayesian optimization can be an interesting direction.

Second, decision-level integration allows independent analysis of each data set, while the integration at this upper level also limits the possible interactions among different data modalities. Using the knowledge and decision from a resource-rich modality to assist decision making in another modality is a challenging but rewarding task. For example, we often only have limited data samples (patients' data). However, we have a large collection of biomedical domain knowledge, documented as research papers or medical ontologies. Transferring decision from such knowledge or retrieving such knowledge for decision making [77] can be helpful for physicians.

Third, bridging the huge gap in the decision-level integration can improve the final clinical action-taking. After we design accurate predictive models with these integration methods, suggesting viable actions (medications and procedures) for physicians remains a challenge. Because even though randomized trials are widely used in drug and clinical trials to determine the effect of new technology, in daily practice, physicians can only resort to observational study for such evaluations [78]. To solve the problem, we need to understand the causal relationship between genomic data, medical conditions, and treatment to suggest and evaluate more reasonable actions. However, causal inference modeling is a huge topic in biomedical big data analytics on its own, and we will not discuss it further in this dissertation.

2.2 Multi-modal data integration by consensus and complementary principles

The advancement of biomedical techniques such as next-generation sequencing (NGS) and wearable devices have generated high-throughput multi-modality data enabling a more comprehensive view of the patients for personalized and precision care. Besides lab tests

and medical imaging, physicians can order genetic tests to obtain patient genomics information to make more accurate diagnoses and decisions. However, how to utilize these multi-modal data remains a major challenge for both researchers and clinicians.

Multi-modal data analysis, including feature selection and predictive modeling, is more complex than single-modality analysis due to the cross-modality information. There are three ways to address this: to consider the cross-modality interaction and identify correlated features among modalities in feature selection stage [79]; to integrate the multi-modality information at the intermediate feature stage [80, 81, 82]; or to integrate at the decision stage [83]. Although substantial progress has been made in this area, modeling the interactions among modalities and getting rid of redundant or irrelevant information remains extremely challenging.

2.2.1 Multi-view/ Multi-modal Learning

Data integration algorithm investigates how to combine multi-modal features of the same patient to achieve improved prediction performance, also known as multi-view learning. Multi-view/ multi-modal learning [84, 85] are machine learning techniques for building models that can integrate multiple types of input information (called 'views' or 'modalities' in the literature). For example, in the biomedical field, a patient's electronic health records (EHRs), personal health records (PHRs), genomic and proteomic variations, and medical images can be considered various views. Ideally, these multi-view data should be jointly evaluated to realize a more personalized diagnosis and treatment.

The most naive approach is to treat the multi-view learning problem as a single-view problem, where all views are concatenated into a single view and then solved with the established single-view models. However, as the feature concatenation ignores the interactions between views and the number of features increases as the number of modalities increases, this naive multi-view method's performance can be sub-optimal with issues such as over-fitting.

To exploit the multi-view data, we have to consider the interactions among modalities. There are two principles in multi-view learning [84]: 1) the consensus principle, which assumes that the disagreement between views upper bounds the classification errors. Thus, we should aim to maximize the agreement between views. 2) the complementary principle, which assumes that each view contains information other views do not have, and we should extract the difference from each view while preserving the shared information. Researchers have developed two approaches, model-agnostic and model-based, based on these two principles to combine data from multiple modalities.

Model-agnostic approaches are simple in design and usually utilize the complementary principle. Based on when data from different modalities are integrated, we have early integration (where we concatenate the raw/pre-processed features), late integration (where we combine the output from each learning algorithm), and hybrid integration (where we use early and late integration together).

Model-based approaches design models for integrating different modalities: 1) kernel-based algorithms, mainly multiple kernel learning (MKL) [63], first compute kernel matrices for each modality, then combines kernels in a linear or non-linear fashion for succeeding kernel-based classification or regression algorithms. As kernels evaluate the similarities between data points, using modality-specific kernels helps capture heterogeneous information from each modal and improves the performance. Kernel methods are especially helpful for small sample sizes but suffer from high computational complexity when sample sizes are large. 2) graphical models such as Bayesian networks and Markov random fields are a class of algorithms that treat each feature as a random variable and exploit the probability relationships among them [86]. This approach's benefit is that we can incorporate more priors into our modeling and easily interpret the models. However, graphical models are also computationally expensive. 3) deep learning-based multi-modal learning algorithms [87] gain increasing popularity in the literature for the past few years. We design a modality-specific neural network for feature extraction for each modality, and the extracted features

can be fused for downstream analysis. This approach's benefit is that deep neural networks excel in extracting non-linear features and can easily incorporate additional regularization. Popular architectures for deep learning-based multi-modal integration include joint representation, coordinated representation, and cross-modality autoencoders [88]. The whole architecture can be trained end-to-end with gradient-based optimization algorithms, making the approach scalable to large sample scenarios.

2.2.2 Multi-modal Learning in Biomedical Science

With the development of data collection technologies in the biomedical domain, researchers now have access to various multi-modal data such as high-throughput multi-omics data (e.g., gene expression, DNA methylation, and single nucleotide polymorphisms (SNPs)), medical imaging data (e.g., pathology and radiology), and clinical records (e.g., demographics, insurance claims, and past medication history).

Based on characteristics of the modality available, researches have designed various task-specific multi-modal learning algorithms in recent biomedical research [89, 90]. In the sub-filed of -omics data analysis, various multi-modal learning methods have been proposed for multi-omics integration. For example, Xu et al. [91] studied how to integrate -omics data using graph-based similarity between molecular information such as gene expression and DNA methylation and showed superior performance in cancer types classification and survival prediction. Vasaikar et al. [92] created a database of over one billion data points by combining multi-omics data and clinical data from the TCGA dataset for 32 cancer types with the proteomics data. In addition, they presented a module for analyzing and visualizing associations between clinical and molecular attributes. Way et al. [93] integrated RNA-seq, copy number variations, and mutations for identification of abnormal molecular states in tumors. Ma et al. [81] studied how to integrate multi-omics data using neural network-based approaches by combining multiple autoencoders and how domain knowledge can be incorporated to improve the learned representations. The improved

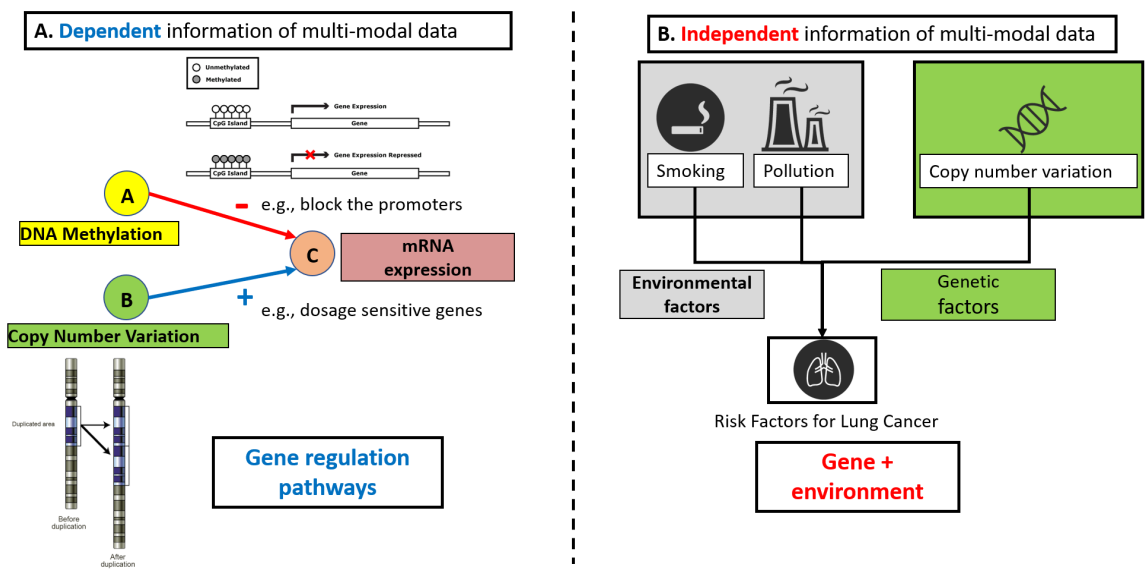


Figure 2.2: Two application scenarios for consensus and complementary principles. A. For dependent modalities such as multi-omics, we can apply the consensus principle for integration. B. For independent modalities such as gene+environmental data, we can apply the complementary principle for integration.

representations outperform other integration methods in predicting disease progression on TCGA datasets. A similar autoencoder-based approach has been proposed by Chaudhary et al. [94], where they used k-means clustering to identify survival-risk subgroups and showed that integration of mRNA, miRNA, and DNA methylation improved the survival prediction for cancer patients. Huang et al. [82] integrated intermediate representations from multi-layer perceptrons and showed improved performance on cancer survival prediction on select TCGA datasets.

Based on multi-view learning's complementary and consensus principles, we have identified two application scenarios for biomedical multi-modal data integration: dependent and independent modalities. For dependent modalities that are closely related to each other (e.g., multi-omics) through association/causal relationships, we propose to capture the interactions among modalities through the consensus principle. For independent modalities that are less connected to each other (e.g., various environmental and lifestyle factors), we propose to combine the multi-modalities through the complementary principle (Figure 2.2).

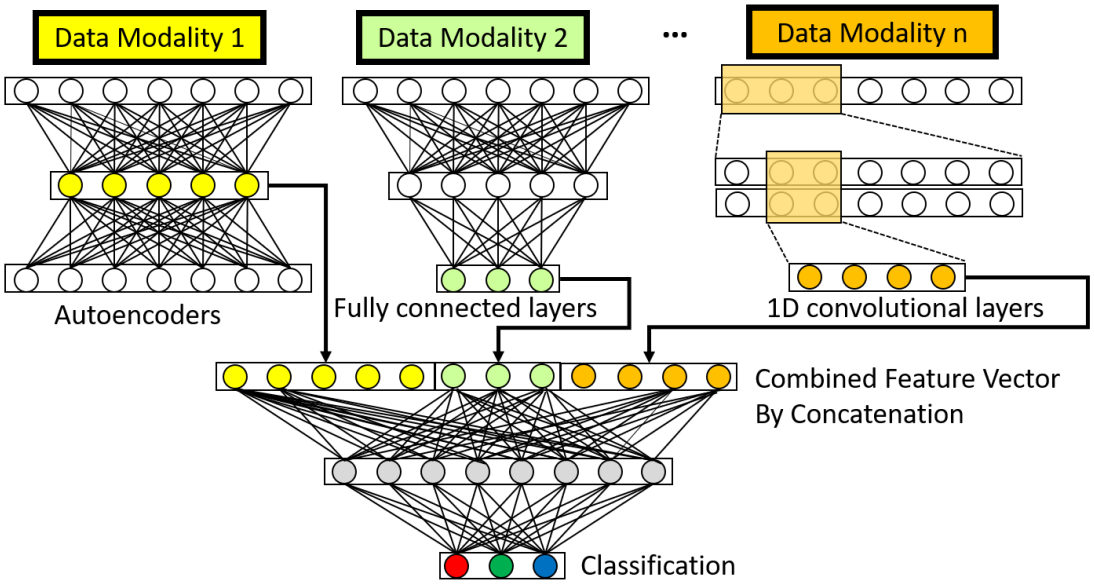


Figure 2.3: Integrating independent multi-modal biomedical data with complementary principle. We can apply modality-specific deep networks for each data modality independently and then combine the hidden features learned for each modality by concatenation.

To integrate independent multi-modal biomedical data with complementary principle, we propose a straightforward framework by first applying modality-independent feature representations and then combining the hidden features learned for each modality by concatenation (Figure 2.3). We have applied the proposed intermediate feature-level concatenation-based integration methods to the prediction of Alzheimer's disease in the following section.

2.3 Multi-modal data integration for early detection of Alzheimer's disease

2.3.1 Background

Alzheimer's disease (AD) is the most common neurodegenerative disorder and forms the sixth leading cause of death in the United States. The worldwide disease burden of AD is predicted to reach $2 trillion by 2030. Thus, the early detection of AD is essential to reduce the cost while improving healthcare quality. Despite extensive research and advances in clinical practice, less than 50% of the people with AD are being diagnosed accurately for their pathology and disease progression based on their clinical symptoms. The most

conclusive evidence for Alzheimer's diagnosis is the presence of amyloid plaques and neurofibrillary tangles in histopathology. However, the early onset of AD is not correlated with plaque presence, but with synaptic and neuronal loss [95].

Research on data from Alzheimer's disease initiative and data mining strategies for AD [96, 97, 98] are being undertaken to improve our understanding of the underlying disease processes. AD biomarkers including clinical symptoms [99] (e.g., dementia and memory loss) and neurological test scores (e.g., MMSE scores) are being augmented with imaging, genetic, and protein-related biomarkers [100, 101, 102, 103]. Most of these studies identify biomarkers using a single-modality data, which restricts holistic assessment of AD disease progression. Thus, there have been AD multi-modal analyses that combine various imaging modalities [104, 105, 106, 107, 108, 109] such as structural MRI (T1 weighted, T2 weighted), fMRI and PET [110, 111], and imaging genetics [112]. In addition, genetics have been used with clinical data to augment data labels and phenotypes. Besides shallow learners, DL models such as auto-encoders [113] and deep-belief networks [114] have been used for PET and MRI image data fusion with improved prediction.

In this case study, we further the multi-modal AD data fusion to advance AD stage prediction by using DL to combine imaging, EHR, and SNP data to classify patients into control, MCI, and AD groups. We use stacked denoising autoencoders for EHR and SNP data feature extraction and novel 3D CNNs for MRI imaging data. After the networks are separately trained for each data modality, we combine them using different classification layers, including decision trees, random forests, support vector machines (SVM), and k-nearest neighbors (kNN). We demonstrate the performance of our integration models using the Alzheimer's Disease Neuroimaging Initiative (ADNI) [115] dataset that contains SNP (808 patients), MR imaging (503 patients), and clinical and neurological test data (2,004 patients). The proposed intermediate-feature-level integration using novel DL architectures outperform shallow learning models [116].

2.3.2 Methods

Dataset

The data we use in this case study are obtained from the ADNI database (adni.loni.usc.edu) [115]. The primary goal of ADNI is to measure the progression of mild cognitive impairment (MCI) and early AD by combining serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments. The ADNI data repository contains imaging, clinical and genetic data for over 2,220 patients spanning over four studies (ADNI1, ADNI2, ADNI GO, and ADNI3). Our study focuses on ADNI1, 2, and GO because ADNI 3 is an ongoing study due to end in 2022. The data is currently being released in phases with limited availability for imaging (unprocessed) and no genetic data yet. The imaging data (ADNI1, 2, and GO) consists of MRI and PET images, of which we use cross-sectional MRI data corresponding to the baseline screenings from ADNI1 (503 patients). The data publisher has standardized the images to eliminate the non-linearities caused by the scanners from different vendors. For clinical or EHR data, we use 2,004 patients (ADNI1, ADNI2, and ADNI GO) data from the clinical tests (e.g., memory tests, balance tests, and cognitive tests), medication data (e.g., usage of levodopa), imaging score summaries (e.g., levels of FDG from PET, brain volumes from MRI), patient demographics (e.g., age and gender), and biochemical tests. The genetic data consists of the whole genome sequencing (WGS) data from 808 ADNI participants (at the time of sequencing, 128 with AD, 415 with MCI, and 267 controls), sequenced by Illumina's non-CLIA laboratory at roughly 30-40x coverage in 2012 and 2013. The resulting variant call files (VCFs) have been generated by ADNI using Broad best practices (BWA and GATK-haplotype caller) in 2014. In this study, we use a total of 2,004 patients for whom clinical data were available, 503 patients with imaging data (9,108 voxels per patient distributed over 18 slices, with each slice having 22×23 voxels), and 808 patients with genetic data (Table 2.1). For participants with multiple visits,

Table 2.1: A summary of the ADNI data

	CN	MIC	AD	Example Data Types/Features
Clinical Data	598	699	707	Demographics, neurological exams, cognitive assessments, bio-markers (e.g. alanine, choline), medication (e.g. levodopa), imaging summary scores (e.g. brain are volumes)
Imaging Data	132	104	266	Cross-sectional MRI data
Genetic Data	245	338	226	SNPs obtained from whole genome sequencing (WGS) data

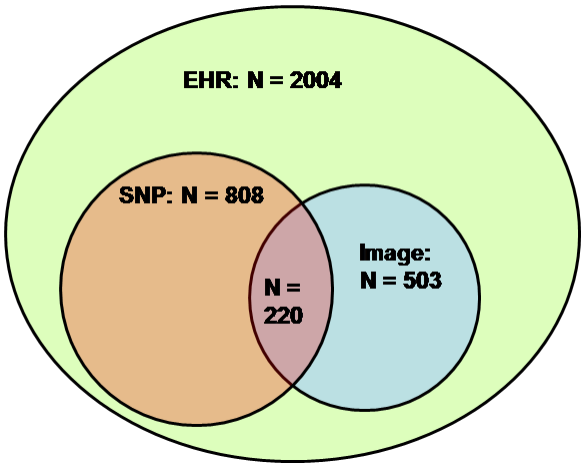


Figure 2.4: Venn diagram of the ADNI data. 220 patients had all the three data modalities, 588 patients had SNP and EHR, 283 patients had imaging and EHR, the remaining patients had only EHR data.

we use the diagnosis from the patient's last visit. As shown in Figure 2.4., 220 patients have all three data modalities, 588 patients have SNP and EHR, 283 patients have imaging and EHR, the remaining patients have only EHR data

Multi-modal data integration

We use the multi-modal data with MR imaging (503 patients), SNPs (808 patients), and the EHR (2004 patients) to predict AD stages. We first demonstrate the superiority of deep models over shallow models such as kNN, one-vs-one coding SVM, random forests, and decision trees for each single data modality. The SNP and EHR features for shallow models and DL are the same. For imaging, when using DL, we apply multi-slice 3D voxels

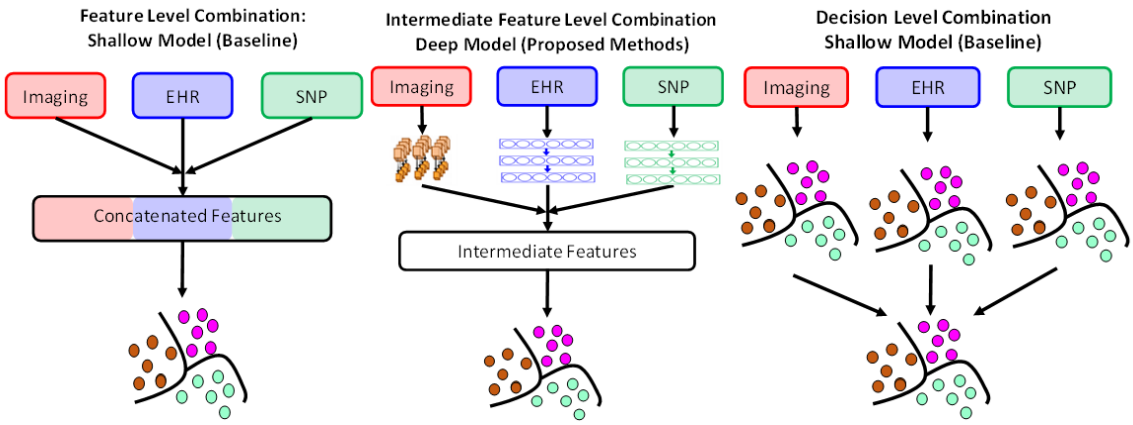


Figure 2.5: Deep Model for Data Integration Compared with Shallow Models of Data Integration. a) Feature level integration on shallow models, where the features are concatenated before passing into shallow models. b) Deep intermediate feature level integration where the original features are transformed separatelyusing deep models prior to integration and prediction. c) Decision level integration where voting is performed using decisions of individual classifiers. In this study, we compare the performance of deep intermediate level integration against shallow feature and decision levels integrations for the prediction of Alzheimer's stages.

directly, while for shallow learners, we extract expert crafted features derived from the 3D voxels.

Regarding AD staging, only EHR has three-stage classes CN, AD, and MCI. SNP expression does not vary between MCI and AD [117], and only has CN vs. AD/MCI prediction. On images, patients with early MCI were structurally similar to CN, and those from patients with late MCI were structurally similar to AD [118]. Thus, only controls (CN) and Alzheimer's disease (AD) are used for staging assessment for imaging data. Thus, combining all three modalities can help us significantly improve AD staging prediction accuracy.

We have developed three data fusion strategies: 1) Feature-level combinations using shallow models, 2) Intermediate-feature-level combinations using deep models, and 3) Decision-level combinations using shallow models (Figure 2.5).

Feature-level combinations are performed through a direct concatenation of the data modalities using shallow learners (Figure 2.5). The intermediate-feature-level combination

is performed by extracting intermediate features using DL, which is then concatenated and passed through a classification layer. Decision-level combinations are performed through voting on the single-modalities. We test shallow models such as kNN, one-vs-one coding SVM, random forests, and decision trees for decision-level combinations and present the best performing model. For the intermediate-feature-level models (Figure 2.6), we evaluate four combinations: 1) EHR + imaging + SNP, 2) EHR + imaging, 3) EHR+ SNP, and 4) imaging + SNP. For all combinations except imaging + SNP, we perform three-stage classification (CN, AD, and MCI). For imaging + SNP, we perform classification into AD vs. CN.

All cases mentioned above are evaluated using internal cross-validation and an external test set. We first remove 10% of the data as an external test set. On the remaining 90%, we perform 10-fold cross-validation with 81% of the total data being used for training and 9% for internal cross-validation. The internal cross-validation data set is used to optimize the model.

2.3.3 Results

We report the ADNI results for both the internal cross-validation partition and the external test dataset. For each of the DL models and the shallow models as baselines, we use mean values of accuracy, precision, recall, and meanF1 scores as metrics to show the superiority of deep models for single-modalities and the improvements gained from data integration.

3D Convolutional Neural Network (DL) is Superior to Shallow Models on Imaging MRI Data

One patient's imaging data consists of 9,108 3D voxels of dimension $22\times23\times18$, corresponding to each of the five selected brain areas. The number of nodes in DL models for the first-level fully connected layers is $5 \times 20 = 100$, and the number of nodes for the second level fully connected layer is 20. The results (Figure 2.7a.) indicate that the CNN

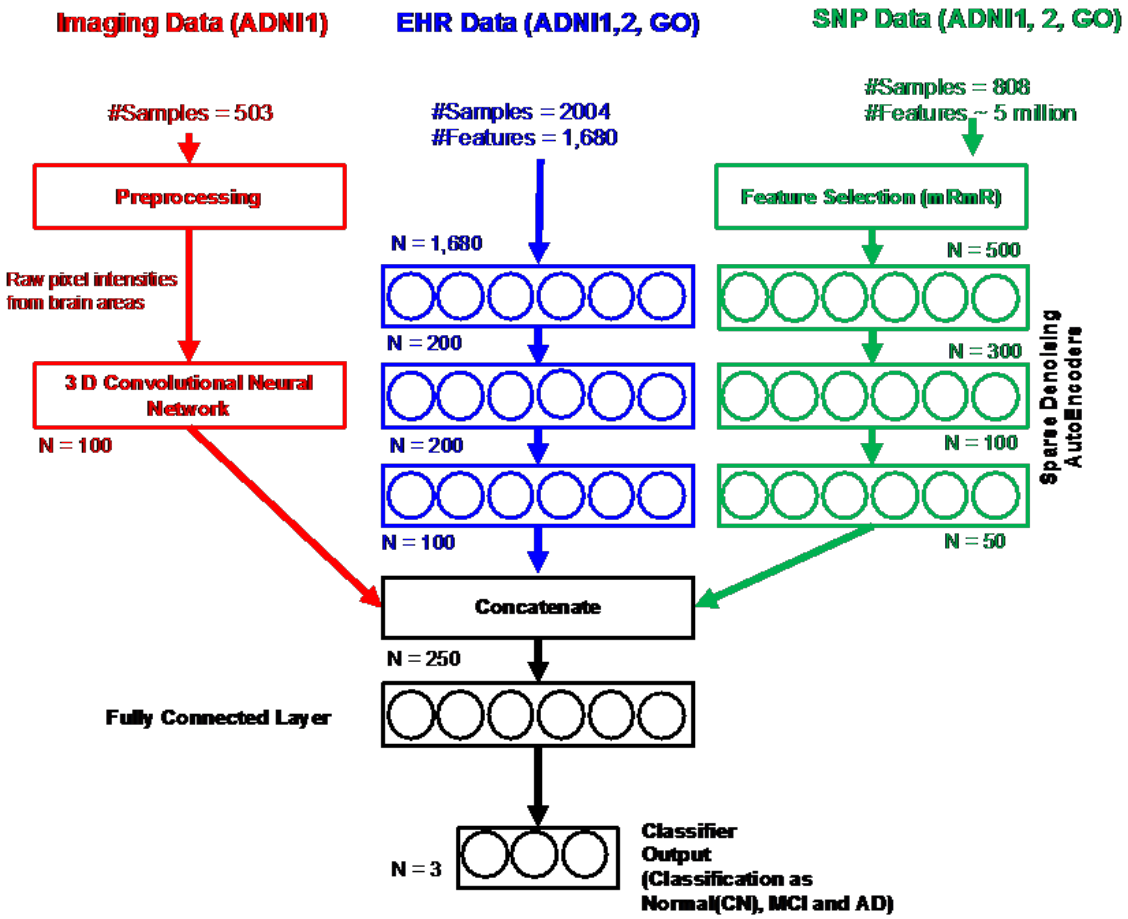


Figure 2.6: Intermediate-Feature-Level Combination Deep Models for Multimodality Data Integration for Clinical Decision Support. Data from diverse sources, imaging, EHR and SNP are combined using novel deep architectures. 3D convolutional neural network architectures used on 3D MR image regions to obtain intermediate imaging features. Deep stacked denoising autoencoders are used to obtain intermediate EHR features. Deep stacked denoising autoencoders are used obtain intermediate SNP features. The 3 types of intermediate features are passed into a classification layer for classification into Alzheimer's stages (CN, MCI and AD).

based imaging models outperform shallow models and give the best precision and meanF1 scores.

Deep Autoencoder Model is Comparable to Shallow Models on EHR Data

EHR data consists of 2,004 patients with 1,680 normalized features per patient, which we use to classify the patients into AD, MCI, and controls (three class). We use a three-layer auto-encoder with 200, 100, and 50 nodes each. The deep networks are trained using Adam with a max epoch count (repetition of DL network training on the entire dataset to allow adequate training) of 25. After hyperparameter optimization, the regularization coefficients for initial training is fixed at 0.03 and those for fine-tuning at 0.03. The dropout probability is set to 0.6 for all the layers. The results (Figure 2.7b.) indicate that the autoencoders outperform shallow models such as kNN and SVM, and they are comparable to decision trees and random forests.

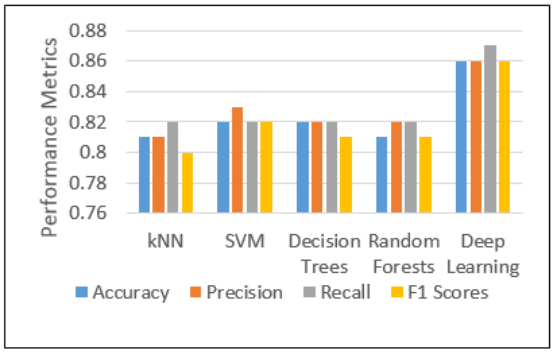
Deep Autoencoder Model is Superior to Shallow Models for SNP Data

Processed SNP data consists of 808 patients with 500 features (each with levels 1,2,3), which we use to classify them into AD/MCI vs. controls (two-class). Autoencoder network consists of three hidden layers with 200, 100, and 50 nodes each. Using Adam optimization and a max epoch count of 30, the best performing models have regularization coefficients for initial training as 0.03 and those for fine-tuning at 0.06. The corruption (dropouts) is 0.6 for each layer. The results (Figure 2.7c.) indicate that the autoencoder models outperform all the baseline models.

Results for Multi-Modality Classification

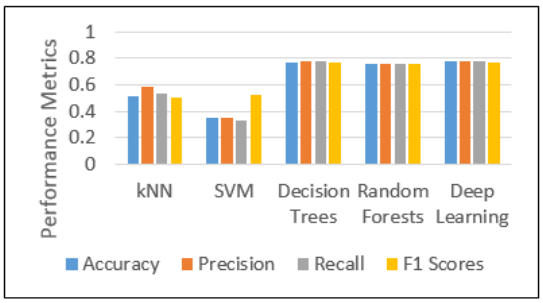
The intermediate features generated from the single-modality deep-models are concatenated and passed to an additional classification layer for integration.

Metrics		kNN	SVM	Decision Trees	RF	Deep Model
Accuracy	CN vs AD	0.81 ± 0.05	0.82 ± 0.09	0.82 ± 0.06	0.81 ± 0.08	**0.86 ± 0.04**
Precision	CN	0.78 ±0.12	0.82 ±0.09	0.82 ±0.14	0.81 ±0.1	**0.92 ±0.08**
	AD	**0.85 ±0.11**	0.84 ±0.13	0.82 ±0.1	0.82 ±0.13	0.80 ±0.1
Recall	CN	0.83 ±0.14	0.81 ±0.15	0.79 ±0.11	0.80 ±0.14	**0.85 ±0.08**
	AD	0.80 ±0.1	0.84 ±0.12	0.85 ±0.09	0.84 ±0.11	**0.89 ±0.1**
MeanF1	CN	0.79 ±0.06	0.80 ±0.1	0.79 ±0.08	0.80 ±0.07	**0.88 ±0.04**
	AD	0.81 ±0.05	0.83 ±0.1	0.83 ±0.06	0.82 ±0.09	**0.84 ±0.07**



a

Metrics		kNN	SVM	Decision Trees	RF	Deep Model
Accuracy	CN	0.67 ± 0.04	0.84 ± 0.03	**0.9 ± 0.02**	0.88 ± 0.03	0.83 ± 0.07
	MCI	0.65 ± 0.03	0.73 ± 0.02	**0.79 ± 0.02**	0.76 ± 0.04	0.74 ± 0.06
	AD	0.78 ± 0.04	0.81 ± 0.02	0.82 ± 0.01	0.83 ± 0.03	**0.85 ± 0.03**
Precision	CN	0.51 ±0.03	0.77 ±0.04	**0.84 ±0.04**	0.81 ±0.05	0.75 ±0.12
	MCI	0.56 ±0.06	0.61 ±0.03	**0.76 ±0.02**	0.67 ±0.05	0.65 ±0.09
	AD	**0.86 ±0.07**	0.77 ±0.03	0.73 ±0.02	0.79 ±0.05	0.84 ±0.07
Recall	CN	0.88 ±0.09	0.77 ±0.08	**0.91 ±0.03**	0.84 ±0.05	0.76 ±0.27
	MCI	0.36 ±0.07	0.64 ±0.05	0.58 ±0.07	**0.66 ±0.07**	0.65 ±0.12
	AD	0.61 ±0.08	0.74 ±0.05	**0.84 ±0.05**	0.77 ±0.05	0.79 ±0.05
MeanF1	CN	0.64 ±0.03	0.77 ±0.05	**0.87 ±0.03**	0.82 ±0.04	0.72 ±0.23
	MCI	0.44 ±0.06	0.62 ±0.03	**0.66 ±0.05**	0.66 ±0.05	0.64 ±0.05
	AD	0.71 ±0.06	0.75 ±0.03	0.78 ±0.02	0.78 ±0.04	**0.81 ±0.04**



b

Metrics		kNN	SVM	Decision Trees	RF	Deep Model
Accuracy	CN vs AD/MCI	0.68 ± 0.04	0.72 ± 0.07	0.69 ± 0.06	0.7 ± 0.04	**0.89 ± 0.03**
Precision	CN	0.51 ±0.27	0.53 ±0.1	0.50 ±0.2	0.48 ±0.13	**0.90 ±0.11**
	AD/MCI	0.73 ±0.04	0.81 ±0.06	0.73 ±0.03	0.75 ±0.05	**0.89 ±0.05**
Recall	CN	0.24 ±0.09	0.57 ±0.1	0.17 ±0.09	0.31 ±0.15	**0.72 ±0.11**
	AD/MCI	0.87 ±0.08	0.78 ±0.07	0.91 ±0.1	0.87 ±0.05	**0.96 ±0.05**
MeanF1	CN	0.29 ±0.08	0.54 ±0.09	0.24 ±0.1	0.36 ±0.13	**0.79 ±0.05**
	AD/MCI	0.79 ±0.03	0.79 ±0.05	0.80 ±0.05	0.80 ±0.03	**0.92 ±0.02**

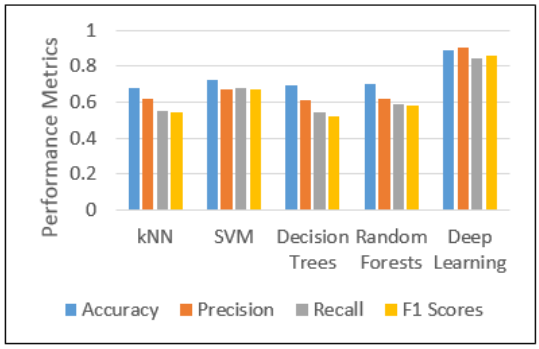
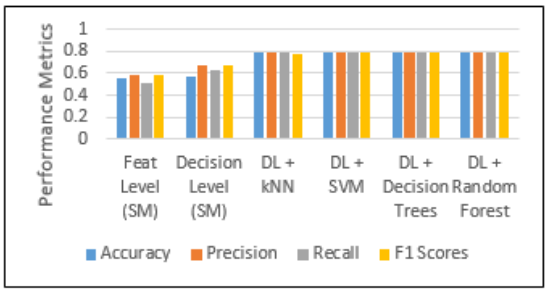


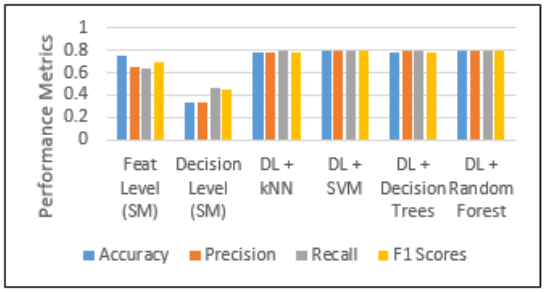
Figure 2.7: Internal Cross Validation Results for Individual Data Modality to Predict Alzheimer's Stage a) Imaging Results: Deep learning prediction performs better than shallow learning predictions b) EHR Results: Deep learning outperforms shallow models kNN and SVM and is comparable to decision trees and random forests c) SNP Results: Deep learning outperforms shallow models. The kNN, SVM , RF and decision trees are shallow models. ((kNN: k-Nearest Neighbors, SVM: Support Vector Machines, and RF: Random Forests).

Metrics		Feature Level (SM)	Decision Level (SM)	DL + kNN	DL + SVM	DL + Decision Trees	DL + RF
Accuracy	CN	0.73 ± 0.13	0.67 ± 0.13	0.87 ± 0.02	0.87 ± 0.02	0.87 ± 0.03	**0.88 ± 0.02**
	MCI	0.57 ± 0.12	0.7 ± 0.12	0.79 ± 0.03	0.79 ± 0.02	0.79 ± 0.04	**0.8 ± 0.02**
	AD	0.61 ± 0.13	0.64 ± 0.09	0.87 ± 0.02	**0.87 ± 0.02**	0.87 ± 0.03	**0.87 ± 0.02**
Precision	CN	0.64 ±0.16	0.51 ±0.17	0.76 ±0.05	0.79 ±0.02	0.79 ±0.05	**0.81 ±0.05**
	MCI	0.24 ±0.17	0.56 ±0.2	0.72 ±0.05	0.70 ±0.04	0.70 ±0.06	**0.72 ±0.03**
	AD	0.62 ±0.14	**1 ±0**	0.87 ±0.04	0.87 ±0.04	0.87 ±0.05	0.86 ±0.04
Recall	CN	0.70 ±0.21	**0.93 ±0.09**	0.9 ±0.04	0.84 ±0.05	0.85 ±0.06	0.85 ±0.06
	MCI	0.22 ±0.16	0.70 ±0.2	0.68 ±0.07	**0.71 ±0.05**	0.71 ±0.08	0.71 ±0.07
	AD	0.62 ±0.24	0.27 ±0.17	0.78 ±0.06	0.8 ±0.03	0.79 ±0.04	**0.81 ±0.05**
MeanF1	CN	0.66 ±0.16	0.65 ±0.14	0.82 ±0.03	0.81 ±0.03	0.82 ±0.04	**0.83 ±0.04**
	MCI	0.26 ±0.09	0.6 ±0.14	0.69 ±0.05	0.70 ±0.04	0.70 ±0.05	**0.71 ±0.03**
	AD	0.61 ±0.17	0.45 ±0.17	0.82 ±0.03	0.83 ±0.02	0.82 ±0.03	**0.83 ±0.03**



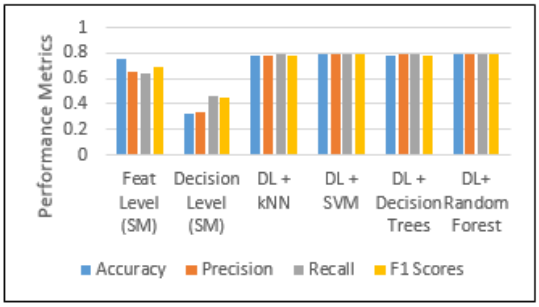
a

Metrics		Feature Level (SM)	Decision Level (SM)	DL + kNN	DL + SVM	DL + Decision Trees	DL + RF
Accuracy	CN	0.87 ± 0.03	0.77 ± 0.04	**0.88 ± 0.02**	**0.88 ± 0.02**	0.87 ± 0.03	**0.88 ± 0.02**
	MCI	0.77 ± 0.06	0.76 ± 0.06	0.79 ± 0.03	**0.79 ± 0.01**	0.79 ± 0.04	0.79 ± 0.02
	AD	0.82 ± 0.04	0.78 ± 0.04	0.86 ± 0.02	**0.87 ± 0.01**	0.87 ± 0.03	0.87 ± 0.02
Precision	CN	**0.81 ±0.08**	0.59 ±0.05	0.79 ±0.03	0.79 ±0.04	0.79 ±0.05	0.79 ±0.04
	MCI	0.72 ±0.07	0.74 ±0.08	**0.72 ±0.06**	0.71 ±0.03	0.7 ±0.06	0.71 ±0.04
	AD	0.72 ±0.09	**1 ±0**	0.85 ±0.05	0.87 ±0.03	0.87 ±0.05	0.85 ±0.03
Recall	CN	0.78 ±0.06	**1 ±0**	0.87 ±0.05	0.88 ±0.04	0.85 ±0.07	0.87 ±0.03
	MCI	0.75 ±0.1	0.74 ±0.06	0.68 ±0.08	0.7 ±0.06	**0.71 ±0.07**	0.69 ±0.06
	AD	0.7 ±0.1	0.31 ±0.12	0.8 ±0.07	**0.79 ±0.04**	0.79 ±0.05	0.8 ±0.05
MeanF1	CN	0.8 ±0.06	0.74 ±0.04	0.83 ±0.03	**0.83 ±0.02**	0.82 ±0.05	0.83 ±0.03
	MCI	0.73 ±0.07	**0.74 ±0.06**	0.69 ±0.04	0.7 ±0.03	0.7 ±0.05	0.7 ±0.04
	AD	0.71 ±0.07	0.46 ±0.15	0.82 ±0.04	**0.83 ±0.02**	0.83 ±0.03	0.83 ±0.03



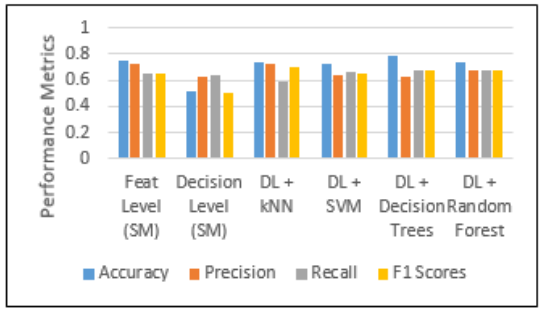
b

Metrics		Feature Level (SM)	Decision Level (SM)	DL + kNN	DL + SVM	DL + Decision Trees	DL + RF
Accuracy	CN	0.85 ± 0.05	0.58 ± 0.09	0.86 ± 0.03	**0.88 ± 0.03**	**0.88 ± 0.03**	0.88 ± 0.04
	MCI	0.78 ± 0.05	0.38 ± 0.1	**0.8 ± 0.03**	**0.8 ± 0.03**	0.79 ± 0.04	**0.8 ± 0.03**
	AD	0.83 ± 0.08	0.38 ± 0.08	0.87 ± 0.03	**0.88 ± 0.02**	0.87 ± 0.03	0.87 ± 0.03
Precision	CN	0.7 ±0.06	0.52 ±0.08	0.75 ±0.04	**0.8 ±0.06**	**0.8 ±0.06**	0.79 ±0.04
	MCI	0.44 ±0.38	0.16 ±0.12	**0.74 ±0.05**	0.71 ±0.05	0.71 ±0.06	0.71 ±0.05
	AD	0.82 ±0.08	0 ±0	0.84 ±0.05	**0.87 ±0.04**	0.86 ±0.06	**0.87 ±0.04**
Recall	CN	0.87 ±0.07	**1 ±0**	0.85 ±0.04	0.86 ±0.04	0.85 ±0.07	0.87 ±0.08
	MCI	0.13 ±0.11	0.38 ±0.19	0.68 ±0.04	**0.71 ±0.08**	**0.71 ±0.08**	0.7 ±0.05
	AD	**0.92 ±0.07**	0 ±0	0.81 ±0.05	0.81 ±0.05	0.8 ±0.05	0.8 ±0.04
MeanF1	CN	0.77 ±0.06	0.68 ±0.07	0.8 ±0.04	0.82 ±0.04	0.82 ±0.05	**0.83 ±0.06**
	MCI	0.27 ±0.1	0.22 ±0.14	**0.71 ±0.04**	0.71 ±0.05	0.71 ±0.06	**0.71 ±0.04**
	AD	**0.86 ±0.07**	0 ±0	0.83 ±0.05	0.84 ±0.03	0.83 ±0.05	0.83 ±0.04



c

Metrics		Feature Level (SM)	Decision Level (SM)	DL + kNN	DL + SVM	DL + Decision Trees	DL + RF
Accuracy	CN vs AD/MCI	**0.75 ± 0.11**	0.51 ± 0.12	0.74 ± 0.08	0.72 ± 0.06	0.73 ± 0.08	0.74 ± 0.09
Precision	CN	0.67 ±0.25	0.37 ±0.12	**0.74 ±0.08**	0.72 ±0.06	0.73 ±0.08	0.74 ±0.09
	AD/MCI	0.78 ±0.12	**0.9 ±0.13**	0.69 ±0.4	0.55 ±0.11	0.54 ±0.12	0.61 ±0.2
Recall	CN	0.39 ±0.15	**0.93 ±0.09**	0.74 ±0.07	0.79 ±0.06	0.81 ±0.05	0.8 ±0.06
	AD/MCI	**0.91 ±0.07**	0.34 ±0.15	0.22 ±0.24	0.49 ±0.18	0.54 ±0.16	0.5 ±0.16
MeanF1	CN	0.48 ±0.15	0.51 ±0.12	**0.97 ±0.05**	0.82 ±0.09	0.8 ±0.09	0.84 ±0.13
	AD/MCI	**0.83 ±0.09**	0.49 ±0.16	0.56 ±0.14	0.5 ±0.12	0.54 ±0.14	0.53 ±0.14



d

Figure 2.8: Internal Cross Validation Results for Integration of Data Modalities to Predict Alzheimer's Stage. a) Imaging + EHR + SNP: Deep learning prediction performs better than shallow learning predictions. b) EHR + SNP: Deep learning prediction performs better than shallow learning predictions. c) Imaging + HER: Deep learning prediction performs better than shallow learning predictions. d) Imaging + SNP: Shallow learning gave a better prediction than deep learning due to small sample sizes. (kNN: k-Nearest Neighbors, SVM: Support Vector Machines, RF: Random Forests, SM: Shallow Models, and DL: Deep Learning).

Combination of all three modalities: (Imaging + EHR + SNP): Deep Model Outperforms Shallow Models

When a particular modality is not available, we mask it as zeros when using DL. The intermediate features from the three modalities are passed to the classification layer. We test kNN, decision trees, random forests, and support vector machines as alternatives for the classification layer. Internal cross-validation (CV) accuracy (Figure 2.8a) using deep models followed by random forests as the classification layer are the best. Deep models for the combination of the three modalities outperform single-modalities DL. In addition, the deep model outperforms shallow models such as feature-level and decision-level for both CV and external test sets during combination.

Combination of SNP and EHR modalities: Deep Model Outperforms Shallow Models

Internal CV accuracy of 0.78 ± 0 using deep models followed by random forests as the classification layer (Figure 2.8b.) is the best. The deep models for EHR + SNP combinations outperform single-modalities DL. The deep model outperforms shallow models such as feature-level combination models for both CV and external test sets during combination.

Combination of Imaging and EHR modalities: Deep Model Outperforms Shallow Models

Internal CV accuracy of 0.79 ± 0 using deep models followed by random forests and SVM as the classification layers (Figure 2.8c.) are the best. The deep models for EHR + imaging combinations outperform single-modalities DL. In addition, the DL model outperforms shallow models such as feature-level and decision-level combination models for both CV and external test sets during combination. Random forests as the classification layer give the best performance on the external set.

Combination of Imaging and SNP modalities: Shallow Model Outperforms Deep Models

We perform two-class classification using a combination of SNP and imaging intermediate features (CN vs. AD/MCI). Internal CV accuracy of 0.75 ± 0.11, using feature-level combination models (Figure 2.8d) is the best. However, the results on the external data are poor. The DL model's performance drop can be attributed to the small overlap of 220 samples between the two modalities. More details of the methods and results can be found in [116].

2.3.4 Concolusion and discussion

For single-modality classification, the DL-based models outperform the shallow models. The hand-crafted features could be the bottleneck for shallow models, which require human expertise and limit the model capability. On the contrary, deep networks can automatically learn optimal feature representations from the training data based on specific objectives. For example, the deep autoencoders learn the feature representation in an unsupervised fashion by reconstructing the input.

Integration of multiple modalities improves the prediction accuracy (three of four scenarios). The deep models for integration also show improved performance over traditional feature-level and decision-level integrations except for the integration of Imaging and SNP data, which are limited by the small number of training data.

One bottleneck for our proposed DL-based data integration performance is the small sample size of the ADNI dataset. To mitigate the sample size challenge, we can utilize strategies such as transfer learning and domain adaptation [118]. For each data modality, we can adopt neural networks pre-trained on other similar datasets (e.g., CNN-based MRI/CT brain imaging classification model trained for other conditions). By composing our model with these pre-trained networks and their parameters, we can perform domain adaptation or fine-tune the network parameters using our labeled ADNI data. On the other hand, we can also perform an unsupervised feature representation learning for each data

modality using publicly available data (e.g., The Cancer Genome Atlas (TCGA) dataset for SNPs). On the other hand, our feature extraction step is performed independently for each modality, which is not trained end-to-end with the integration and classification step. One future direction is to enable end-to-end training for the whole DL-based multi-modal data integration pipeline.

CHAPTER 3

INTEGRATING MULTI-OMICS DATA WITH CONSENSUS LEARNING

In the previous chapter, we have applied intermediate feature-level concatenation to integrate multi-modal data from independent modalities. In this chapter, we will discuss the integration of multi-modal data from dependent modalities. We will focus on integrating multi-omics data (e.g., gene expression and DNA methylation), which are assumed to be connected by associations or causal relationships. We will first present our work on the evaluation of bioinformatics pipelines for gene expression estimation. We will then present two frameworks on the integration of multi-omics data using the consensus principles.

3.1 Impact of RNA-seq Data Analysis Algorithms on Gene Expression Estimation and Downstream Prediction

To use next-generation sequencing technology such as RNA-seq for medical and health applications, choosing proper analysis methods for biomarker identification remains a critical challenge for most users. The US Food and Drug Administration (FDA) has led the Sequencing Quality Control (SEQC) project to conduct a comprehensive investigation of 278 representative RNA-seq data analysis pipelines consisting of 13 sequence mapping, three quantification, and seven normalization methods. In this study, we focused on the impact of the joint effects of RNA-seq pipelines on gene expression estimation as well as the downstream prediction of disease outcomes. First, we developed and applied three metrics (i.e., accuracy, precision, and reliability) to evaluate each pipeline's performance on gene expression estimation quantitatively. We then investigated the correlation between the proposed metrics and the downstream prediction performance using two real-world cancer datasets (i.e., SEQC neuroblastoma dataset and the NIH/NCI TCGA lung adenocarcinoma dataset). We found that RNA-seq pipeline components jointly and significantly impacted

the accuracy of gene expression estimation, and its impact was extended to the downstream prediction of these cancer outcomes. Specifically, RNA-seq pipelines that produced more accurate, precise, and reliable gene expression estimation tended to perform better in the prediction of disease outcome. In the end, we provided scenarios as guidelines for users to use these three metrics to select sensible RNA-seq pipelines for the improved accuracy, precision, and reliability of gene expression estimation, which lead to the improved downstream gene expression-based prediction of disease outcome [119].

3.1.1 Introduction to RNA-seq pipelines evaluation

The first phase of the FDA-led microarray quality control project (MAQC-I) investigated the reliability of microarray platforms for gene expression estimation [52]. The second phase of the project, MAQC-II, studied 30,000+ microarray data analysis pipelines to assess the reproducibility of microarray-based predictive models [120]. Given the rise of the significance of next-generation sequencing in gene expression analysis, the FDA initiated the sequencing quality control project (SEQC) as a continuing MAQC effort to conduct an in-depth assessment of RNA-seq by combining the objectives of both MAQC-1 and MAQC-II [121, 122, 123, 124]. Specifically, the goal of SEQC was to conduct a comprehensive evaluation of both RNA-seq technology and RNA-seq data analysis pipelines, which was similar to the objectives of MAQC-I and MAQC-II for microarrays. While Su et al. summarized the RNA-seq technology investigation [53], this complementary study focuses on the RNA-seq data analysis pipelines targeting medical and health applications. Specifically, this study examines the effect of RNA-seq pipelines on gene expression with three critical metrics (i.e., accuracy, precision, and reliability), and further on downstream gene expression-based prediction of disease outcomes. Although other analyses of RNA-seq are possible (e.g., differential expression analysis [125, 126, 127], alternative splicing [128, 129, 130], and RNA fusion [131, 132]), we focus on gene expression because it is the most widely used genetic variations in biomedical and health applications.

For medical and health applications, choosing a proper RNA-seq gene expression analysis pipeline remains a critical challenge due to its relative immaturity (i.e., fewer standards reported compared with microarrays), complexity, and diverse applicability [133, 134]. We performed a literature survey on RNA-seq pipelines consisting of sequence mapping [135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146], expression quantification [147, 148, 149, 150], and expression normalization [124, 151, 152, 153]. For evaluation of RNA-seq pipelines, the majority strategy is to use some sorts of benchmark datasets, or reference standards to enable quality control [154] (e.g., natural well-characterized genetic materials or synthetic spike-in controls). For benchmark dataset-based evaluation, multiple comparative investigations exist for focusing on individual components of RNA-seq pipelines, such as mapping alone [142, 155, 156, 157, 158, 159, 160], quantification only [150, 130, 161], or normalization [151, 162, 163]. Correspondingly, the joint impact of each of these three components is less understood. For examples, a previous analysis of 50 RNA-seq pipelines examined combinations of ten mapping and five quantification algorithms, but it did not study the effect of different normalization methods' impact and the interaction effect among pipeline components [164]; another study investigated three mapping, two quantification, and five differentially expressed gene (DEG) detection methods, but it did not report the interaction effect among components either [165]; a third study by Sahraeian et al. examined 120 combinations of 39 tools 14 for RNA variant calling, RNA editing, RNA fusion, gene expression, and differential expression, but it did not provide an assessment on their impact on gene-expression-based downstream prediction. Most FDA approvals on medical genomics would be relevant to gene expression applications. To the best of our knowledge, no studies have comprehensively examined the joint effect of RNA-seq pipeline components on gene expression and its downstream prediction of disease outcomes. Thus, this study dedicates to this goal.

The FDA coordinated multiple sites of SEQC to generate a multi-replicate benchmark dataset (referred to as SEQC-benchmark) [53] and a clinical dataset consisting of neurob-

lastoma patient samples (referred to as SEQC-neuroblastoma) [166]. In addition, we have another real-world clinical dataset on lung adenocarcinoma (referred to as TCGA-lung-adenocarcinoma) from The Cancer Genome Atlas (TCGA). These datasets were used to investigate the joint impact of pipeline components on downstream gene expression-based prediction in a two-phase study:

(1) Phase-1: we developed three metrics—accuracy, precision, and reliability—for assessing the performance of a representative set of 278 RNA-seq pipelines (Figure 3.1, blue box) using the SEQC-benchmark dataset (i.e., A, B, C and D where C is 75/25 of A/B while D is 25/75 of A/B). (2) Phase-2: we validated the benchmark metrics by quantifying gene expression in the SEQC-neuroblastoma dataset and the TCGA-lung-adenocarcinoma dataset and demonstrated that the benchmark metrics are informative for inferring downstream prediction of disease outcome (Figure 3.1, pink box).

Our comprehensive investigation revealed that RNA-seq pipeline components—mapping, quantification, and normalization—jointly impacted the accuracy, precision, and reliability of gene expression, and affected the downstream performance of predicting neuroblastoma and lung adenocarcinoma outcome. RNA-seq pipelines that performed well in gene expression estimation also performed well in downstream prediction of disease outcome.

3.1.2 Results

Phase-1: Assessing the joint impact of pipeline components on gene expression estimation with the benchmark metrics

We systematically investigated 278 RNA-seq pipelines that included combinations of mapping, quantification, and normalization components. Sequence mapping algorithms were further categorized based on mapping strategy (i.e., un-spliced and spliced) and mapping reporting (i.e., single-hit and multi-hit) (refer to [119] for more details). To gain insight into these pipelines, we used the SEQC-benchmark dataset and a quantitative PCR (qPCR) benchmark dataset. Because the qPCR results vary among platforms [53], we filtered them

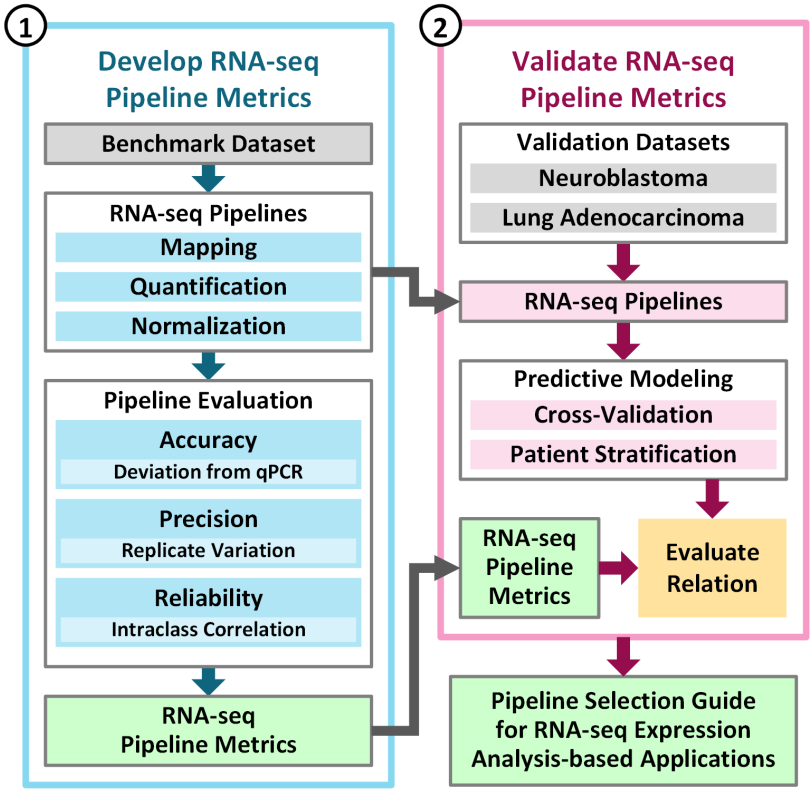


Figure 3.1: The SEQC consortium developed and validated a guideline for selecting RNA-seq pipelines for gene expression-based predictive modeling using the SEQC-benchmark, SEQC-neuroblastoma, and TCGA-lung-adenocarcinoma datasets. Phase-1 of the investigation developed the metrics that captured the accuracy, precision, and reliability of RNA-seq pipelines (the blue box). Using the SEQC-neuroblastoma and TCGA-lung-adenocarcinoma datasets, Phase-2 of the investigation determined that RNA-seq pipeline metrics can be used to select pipelines that result in better performance in terms of predicting cancer outcome (the pink box).

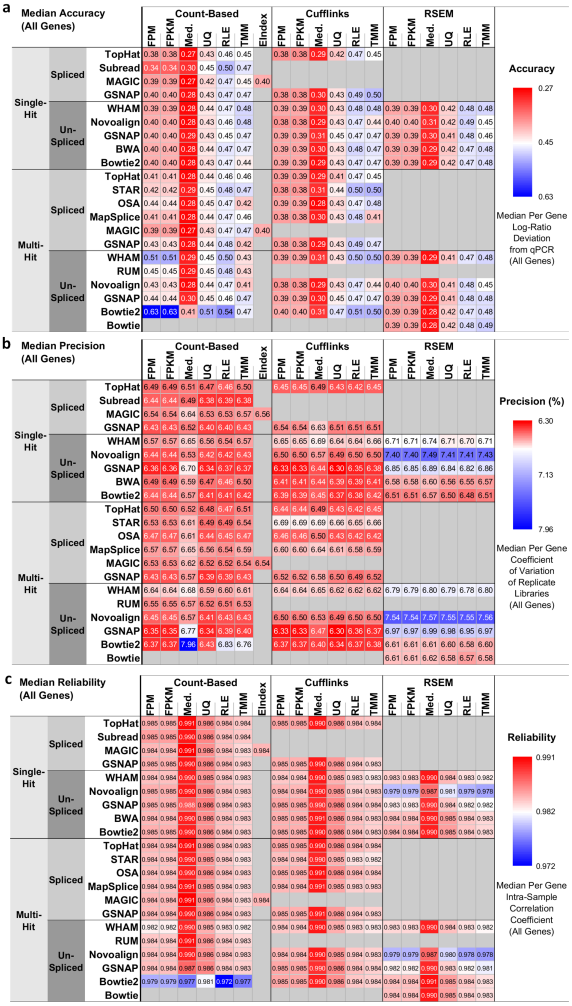


Figure 3.2: The 278 RNA-seq pipelines applied to the SEQC-benchmark dataset differ in terms of gene expression accuracy, precision, and reliability. In each heatmap, the rows are different settings for 13 aligners and the columns are combinations of three quantification and seven normalization methods. (a) Accuracy is defined as the deviation of pipeline-derived log ratios of gene expression from the corresponding qPCR-based log ratios. Median accuracy of all genes (i.e., 10,222 genes) is encoded as color, with red representing the highest accuracy, or the lowest deviation from qPCR. (b) Precision is defined as the coefficient of variation (CoV) of gene expression over replicate libraries. Median precision of all genes is encoded as color, with red indicating the highest precision, or the lowest CoV. (c) Reliability is defined as the intraclass (or intra-sample in our context) correlation that quantifies how similar replicate libraries of a sample are to one another using analysis of variance techniques. Median reliability of all genes is encoded as color, with red representing the highest reliability, or the highest intraclass correlation. Refer to [119] for mathematical definitions of accuracy, precision, and reliability in the context of RNA-seq pipelines.

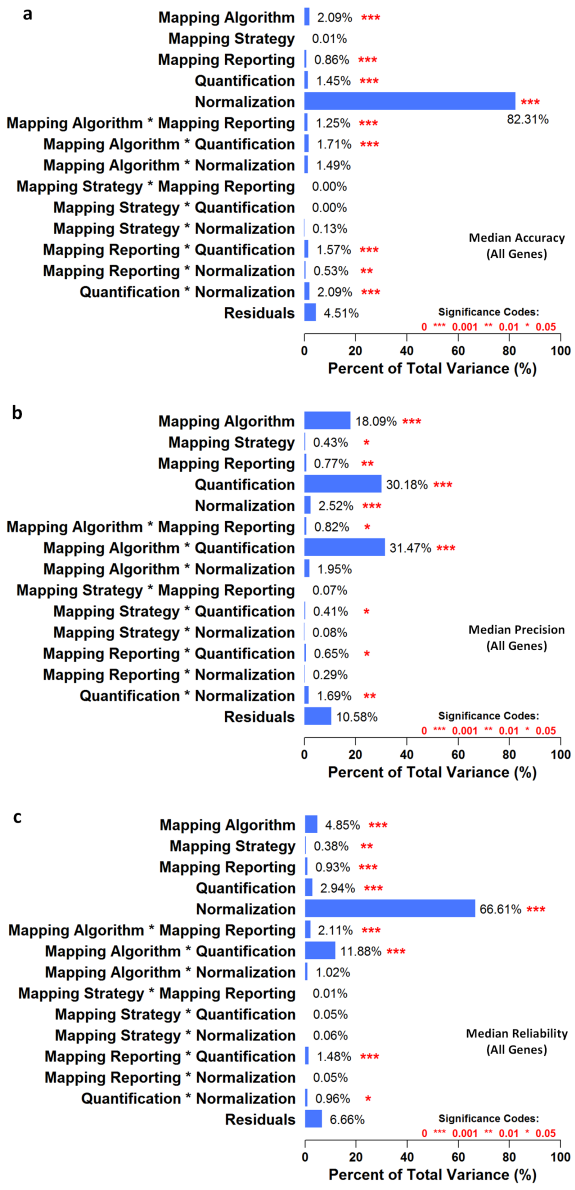


Figure 3.3: Analysis of variance decomposes the overall variance in (a) median accuracy of all genes, (b) median precision of all genes, and (c) median reliability of all genes into various factors considered, including five RNA-seq pipeline components (i.e., mapping algorithm, mapping strategy, mapping reporting, quantification, and normalization) and nine associated two-way interactions. The statistical significance of each component's or interaction's contribution is denoted by red asterisks, with '***' indicating p-values are smaller than 0.001, '**' indicating p-values are smaller than 0.01, and '*' indicating p-values are smaller than 0.05.

to keep the genes that fit the titration ratio of A D samples. After filtering with the titration ratio, we only used 10,222 genes (out of a total of 20,801 genes assayed with qPCR) as a benchmark reference. We have applied three metrics (i.e., accuracy, precision, and reliability) to evaluate each pipeline. These metrics are detailed in the published paper [119]. The joint effect of mapping, quantification, and normalization with respect to these three metrics was assessed comprehensively for all genes (the entire set of 10,222 genes, referred as AllGenes hereafter) and for the low-expression genes (2,044, the subset of AllGenes, referred as LowExpressGenes hereafter).

We defined the accuracy metric as the deviation of RNA-seq pipeline-derived log ratios of gene expression from the corresponding qPCR-based log ratios and visualized the median accuracy of AllGenes and LowExpressGenes using heatmaps (Figure 3.2a). We observed the following results:

(1) Using AllGenes, the log-ratio deviation between RNA-seq and qPCR ranged from 0.27 to 0.63 (Figure 3.2a). A smaller deviation represents higher accuracy. Median normalization exhibited the lowest deviation, or the highest accuracy, compared with all other normalization methods. In addition, for all mapping-quantification combinations, the [Bowtie2 multi-hit + count-based] pipelines showed the largest deviation. Moreover, pipelines with multi-hit mapping and count-based quantification generally showed a larger deviation than other pipelines. Among all pipeline factors, normalization was the largest statistically significant ($p < 0.05$) source of variation (Figure 3.3a). (2) The log-ratio deviation using LowExpressGenes was larger than that using AllGenes, and it ranged from 0.45 to 0.69. The trends of pipeline performance were similar to those using AllGenes, and normalization was also the largest statistically significant (p¡0.05) source of variation. (3) In summary, median normalization with most mapping and quantification algorithms, except for the [Bowtie2 multi-hit + count-based] pipelines, was the best choice for quantifying genes with high accuracy, or low deviation from qPCR.

We defined the precision metric as the coefficient of variation (CoV) of gene expres-

sion across replicate libraries, and visualized the median precision of AllGenes and LowExpressGenes using heatmaps (Figure 3.2b). We observed the following results:

(1) Using AllGenes, the CoV ranged from 6.30% to 7.96% (Figure 3.2b). Smaller CoV represents higher precision. Pipelines with any of Novoalign, GSNAP un-spliced, or WHAM mapping, and RSEM quantification resulted in higher CoV, or lower precision, despite the choice of normalization methods. In addition, the [Bowtie2 multi-hit + count-based + med.] pipeline led to the largest CoV. Moreover, for each mapping-normalization combination, pipelines with either count-based or Cufflinks quantification reported higher precision than those with RSEM quantification, except the [Bowtie2 multi-hit + count-based + med.] pipeline mentioned previously. Quantification, mapping algorithm, and their interaction were the largest statistically significant (p¡0.05) sources of variation (Figure 3.3b). (2) The CoV using LowExpressGenes was larger than that using AllGenes, and it ranged from 11.0% to 15.6%. The trends of pipeline performance were similar to those using AllGenes, except that the [Bowtie2 multi-hit + count-based] pipelines exhibited the highest precision among others. Again, quantification, mapping algorithm, and their interaction were the largest statistically significant (p¡0.05) source of variation. (3) In summary, pipelines with any of Bowtie2 multi-hit, GSNAP un-spliced, or Subread mapping and either count-based or Cufflinks quantification, except for the [Bowtie2 multi-hit + count-based + med.] pipeline, were the best choice for quantifying genes with high precision, or low CoV.

We defined the reliability metric as the intra-class (i.e., intra-sample in the context of the SEQC-benchmark dataset) correlation (ICC) of gene expression, and visualized the median reliability of AllGenes and lowExpressGenes using heatmaps (Figure 3.2c). We observed the following results:

(1) Using AllGenes, the ICC ranged from 0.972 to 0.991 (Figure 3.2c). Larger ICC represents higher reliability. Median normalization exhibited the highest ICC, or the highest reliability, compared with all other normalization methods. In addition, pipelines with Novoalign mapping and RSEM quantification resulted in lower ICC for all but Median

normalization. Moreover, the [Bowtie2 multi-hit + count-based] pipelines showed the lowest ICC. Furthermore, for each mapping-normalization combination, pipelines with either count-based or Cufflinks quantification always reported higher ICC than those with RSEM quantification, except the [Bowtie2 multi-hit + count-based] pipelines mentioned previously. Normalization was the largest statistically significant ($p < 0.05$) source of variation (Figure 3.3c), followed by two-way [mapping algorithm*quantification] interaction. (2) The ICC using LowExpressGenes was smaller than that using AllGenes, and it ranged from 0.938 to 0.975. The trends of pipeline performance were similar to those using AllGenes, except that the [Novoalign + RSEM] pipelines exhibited the lowest ICC, followed by the [Bowtie2 multi-hit + count-based] pipelines. Normalization, two-way [mapping algorithm*quantification] interaction, quantification, and mapping algorithm were the largest statistically significant ($p < 0.05$) sources of variation. (3) In summary, median normalization along with most mapping and quantification algorithms, except for the [Bowtie2 multi-hit + count-based] and [Novoalign + RSEM] pipelines, was the best choice for quantifying genes with high reliability, or high ICC.

We also examined whether the performance of metrics depended on the characteristics of sequence mapping results. We used M-estimation with Huber weighting to fit robust linear models that capture the relationship between the benchmark metrics and alignment profiles. The accuracy metric correlated with the number of mismatches per mapped read, and the precision and reliability metrics correlated with the number of mapped fragments. Fewer mismatches per read and more mapped fragments tended to lead to more accurate, precise, and reliable gene expression.

In summary, the Phase-1 investigation using the SEQC-benchmark dataset demonstrated that gene expression estimation is significantly impacted by the joint effect of multiple RNA-seq pipeline components (Figures 3.2 and 3.3).

Phase-2: The impact of RNA-seq pipeline on the disease outcome prediction performance

We used the SEQC-neuroblastoma and TCGA-lung-adenocarcinoma datasets to assess the impact of upstream RNA-seq pipeline components on the downstream prediction of disease outcome using gene expression (Figure 3.1). The SEQC-neuroblastoma dataset, provided by the SEQC consortium, contains RNA-seq data of 176 primary neuroblastomas obtained from high-risk patients with well-annotated clinical data [166], in which survival information, including event-free survival (EFS) and overall survival (OS), was used for defining group labels. The TCGA-lung-adenocarcinoma dataset contains RNA-seq data of patients with known survival time used for defining group labels.

We used the same set of 278 RNA-seq pipelines to process the SEQC-neuroblastoma dataset (we used only 156 out of the 278 pipelines for the TCGA-lung-adenocarcinoma dataset). For each set of estimated gene expression (278 for neuroblastoma and 156 for lung adenocarcinoma), we performed nested cross-validation using three classifiers—adaptive boosting, logistic regression, and support vector machines, which are proven to be robust and mostly used in machine learning. For each clinical endpoint—neuroblastoma EFS, neuroblastoma OS, and lung adenocarcinoma survival—we calculated the AUC (area under the ROC curve) and MCC (Matthews correlation coefficient), and visualized these using heatmaps. We observed the following results:

(1) For the neuroblastoma EFS endpoint, pipelines using count-based quantification with TMM, RLE, upper quartile, or median normalization tended to achieve high AUC and MCC; while those with FPM or FPKM normalization tended to perform poorly. In addition, Novoalign with Cufflinks and Bowtie2 or BWA with RSEM led to poor AUC and MCC, especially when combining with FPM or FPKM normalization. (2) For the neuroblastoma OS endpoint, median normalization led to higher AUC and MCC than other normalization methods for most mapping-quantification combinations. GSNAP un-spliced mapping performed well with count-based or Cufflinks quantification but not RSEM quantification. In addition, pipelines with RSEM quantification and any of upper quartile, RLE, or TMM

normalization tended to result in poor AUC and MCC. (3) For the lung adenocarcinoma survival endpoint, pipelines with count-based quantification and TMM normalization tended to achieve high AUC and MCC. TopHat alignment with either count-based or Cufflinks quantification also performed well. In contrast, pipelines with any of Novoalign single-hit, STAR, GSNAP un-spliced multi-hit, or Bowtie2 multi-hit and Cufflinks resulted in lower AUC and MCC. (4) ANOVA for each neuroblastoma endpoint showed that normalization was the largest statistically significant ($p < 0.05$) source of variation, followed by mapping algorithm, two-way [mapping algorithm*quantification] interaction, and two-way [quantification*normalization] interaction. For the lung adenocarcinoma endpoint, several pipeline components and their interactions contributed more evenly to the overall variance that may be due to only 156 pipelines were conducted. All ANOVA reported large residual variance that should be explained by higher-order interactions.

These results suggested that the choice of upstream RNA-seq pipeline components significantly impacted the performance of downstream prediction of disease outcome. We summarized the predictive modeling performance for the 278 and 156 RNA-seq pipelines applied to the SEQC-neuroblastoma and TCGA-lung-adenocarcinoma datasets, respectively.

We ranked the 278 RNA-seq pipelines base on the average rank of a combination of the three metrics. The top 10% pipelines were chosen as the good-performing pipelines, while the bottom 10% pipelines were chosen as the poor-performing pipelines. We then compared good-performing versus poor-performing pipelines in conducting gene-expression-based prediction of disease outcome and the success rates of patient stratification for the three endpoints. The comparison was assessed with the one-sided Wilcoxon rank-sum test (Figure 3.4).

For the prediction of neuroblastoma OS endpoint, average prediction performance (i.e., AUC and MCC) of good-performing pipelines was statistically significantly (p¡0.05) larger than that of poor-performing pipelines (Figure 3.4a and Figure 3.4b). For the prediction of

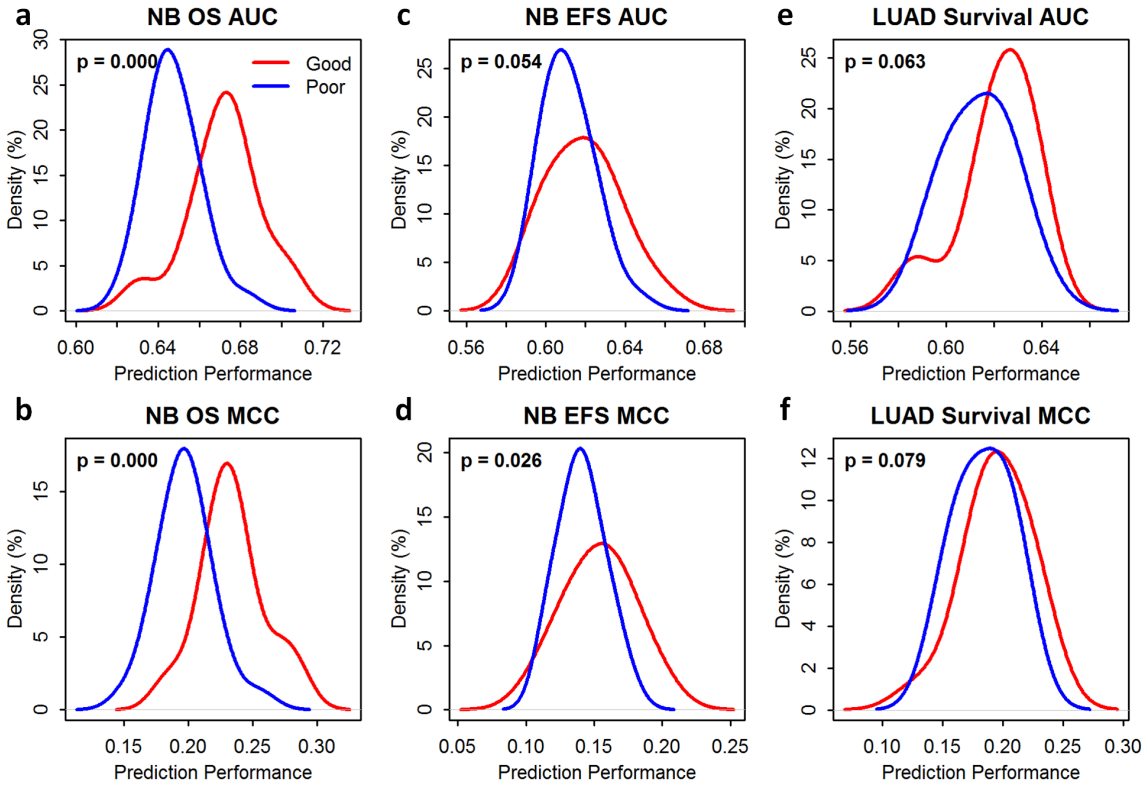


Figure 3.4: RNA-seq pipelines selected based on benchmark metrics (i.e., accuracy, precision, and reliability) were informative for inferring the performance of gene-expression-based prediction of disease outcome—(a) prediction performance measured by the area under the receiver operating characteristic curve (AUROC, or AUC) for the overall survival (OS) endpoint of the SEQC-neuroblastoma (NB) dataset; (b) prediction performance measured by the Matthews correlation coefficient (MCC) for the OS endpoint of the SEQC-NB dataset; (c) prediction performance measured by the AUC for the event-free survival (EFS) endpoint of the SEQC-NB dataset; (d) prediction performance measured by the MCC for the EFS endpoint of the SEQC-NB dataset; (e) prediction performance measured by the AUC for the survival endpoint of the TCGA-lung-adenocarcinoma (LUAD) dataset; and (f) prediction performance measured by the MCC for the survival endpoint of the TCGA-LUAD dataset. The red line in each panel shows the probability density of the prediction performance of good-performing RNA-seq pipelines selected based on benchmark metrics; and the blue line demonstrates that of poor-performing pipelines selected based on the same. Statistical significance (i.e., p-values) was determined using the one-sided Wilcoxon rank-sum test. Panels (a), (b), and (d) show a statistically significant difference (p¡0.05) between the two groups (i.e., the prediction performance of good-performing pipelines vs. that of poor-performing pipelines). The good-performing (Top 10%) and poor-performing pipelines (Bottom 10%) were determined based on the average rank of each RNA-seq pipeline over all benchmark metrics of both all and low-expressing genes.

neuroblastoma EFS endpoint, the average MCC of good-performing pipelines was statistically significantly (p¡0.05) larger than that of poor-performing pipelines (Figure 3.4c), and the average AUC of good-performing pipelines was larger than that of poor-performing pipelines with p slightly larger than 0.05 (Figure 3.4d). For the prediction of LUAD survival endpoint, average prediction performance (i.e., AUC and MCC) of good-performing pipelines was larger than that of poor-performing pipelines (Figure 3.4e and Figure 3.4d) but also with p slightly larger than 0.05.

In addition, good-performing pipelines (e.g., the [GSNAP un-spliced single-hit + Cufflinks + median] pipeline) tended to result in higher success rates of patient stratification than poor-performing pipelines (e.g., the [BWA + RSEM + RLE] pipeline). Figure 3.5 demonstrates Kaplan-Meier estimated survival functions for high-risk and low-risk patients for all endpoints. Good-performing pipelines tended to achieve statistically significant separation (p¡0.05) of the two patient groups (Figure 3.5a-c), while poor-performing pipelines were more likely to fail (Figure 3.5d-f). We have summarized the success rates of patient stratification (i.e., $p < 0.05$ based on the two-tailed log-rank test) for the 278 and 156 RNA-seq pipelines applied to the SEQC-neuroblastoma and TCGA-lung-adenocarcinoma datasets, respectively.

3.1.3 Conclusion and Discussion

We performed a systematic investigation of the 278 representative RNA-seq pipelines in two sequential phases. In Phase-1, we developed three metrics to characterize RNA-seq pipelines using the SEQC-benchmark dataset: 1) accuracy measures gene expression estimation against the qPCR ground truth results, providing justification to support downstream biological interpretation; 2) precision assesses the fluctuation of such a measurement across replicates, estimating the measurement behavior; 3) reliability: the consistency of such a measurement among all the samples, offering the confidence of such a measurement. All these metrics are of great value to support the downstream biological

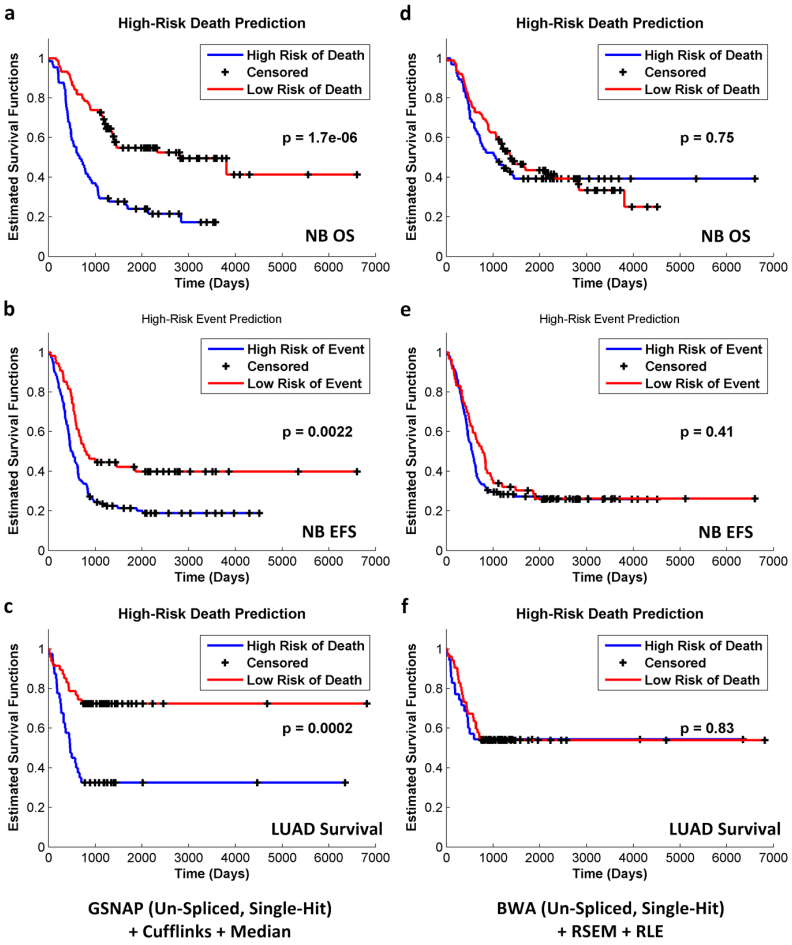


Figure 3.5: The RNA-seq pipeline selection guide was validated by assessing the ability of pipelines to stratify patients based on Kaplan-Meier survival analysis. For each pipeline, patients were grouped by predictive labels (i.e., high risk vs. low risk), and two Kaplan-Meier curves were plotted. The two-tailed log-rank test was used to determine the statistical significance of the separation between the two curves. For good-performing pipelines selected based on benchmark metrics, the success rates of patient stratification (i.e., predictive labels led to a statistically significant separation of Kaplan-Meier curves) were higher. For example, the success rates of the [GSNAP (un-spliced, single-hit) + Cufflinks + Median] pipeline were 93%, 70%, and 67% for the SEQC-NB OS, SEQC-NB-EFS, and TCGA-LUAD-Survival endpoints, respectively. Panels (a) to (c) demonstrate the most statistically significant separation of the two Kaplan-Meier curves for each endpoint. In contrast, poor-performing pipelines led to lower success rates of patient stratification. For instance, the success rates of the [BWA (un-spliced, single-hit) + RSEM + RLE] pipeline were 33%, 30%, and 33% for the SEQC-NB OS, SEQC-NB-EFS, and TCGA-LUAD-Survival endpoints, respectively. Panels (d) to (f) demonstrate the least statistically significant separation of the two Kaplan-Meier curves for each endpoint.

interpretation of gene expression results. We observed that RNA-seq pipeline components jointly affected gene expression estimation (Figures 3.2 and 3.3). This comprehensive investigation provides a framework to assist pipeline selection, which had not previously been reported where individual RNA-seq pipeline components are usually focused (e.g., mapping [159]).

We have summarized and compared the results of our study to previous studies focusing on individual pipeline components. For example, previous studies observed that RUM, GSNAP spliced, STAR, and MapSplice mapping led to more accurate base-level alignment and splice junction detection [142, 156]. In addition, BWA, Bowtie, and Bowtie2 mapping were reported to be robust to sequencing errors and indels [155]. We similarly observed considerable differences in alignment profiles among mapping algorithms, and such differences led to variations in the benchmark metrics. For example, Bowtie2 multi-hit mapping aligned many more reads, a higher percentage of which were sub-optimal mapping variants (i.e., secondary mappings, mismatches, insertions, deletions, and splicing) than WHAM single-hit mapping. Consequently, pipelines with Bowtie2 multi-hit mapping resulted in a larger deviation from the qPCR reference, or lower accuracy, than those with WHAM single-hit mapping. However, such the observation applied to only count-based quantification but not Cufflinks or RSEM (Figure 3.2a). In addition to the observations corresponding to previous literature, we also observed a joint effect between mapping and quantification components.

Variations in mapping performance propagated to the quantification stage. The quantification strategy for multi-hit mappers may explain the variation in gene expression accuracy. For example, Cufflinks and RSEM use Poisson distribution-based models and assign probabilities to each mapping, while HTSeq simply counts total mapped reads regardless of quality. Thus, Cufflinks and RSEM can better handle multi-hit information, resulting in a smaller deviation from the qPCR reference (Figure 3.2a). It is worthwhile to point out that, although variation in performance was observed for different pipelines investigated,

singling out one pipeline for general application is not justified. This is specifically true by giving the fact that the joint effect of multiple components plays a role. Therefore, we aim to demonstrate the importance of pipeline selection and their impact on downstream analysis, such as classification.

In Phase-2, using the SEQC-neuroblastoma and TCGA-lung-adenocarcinoma datasets, we showed that RNA-seq pipelines with better performance in gene expression estimation (using average ranks of three metrics) resulting in a better downstream prediction of disease outcome. Subsequently, we demonstrated the effect of pipeline selection on patient stratification and provided a guideline for pipeline selection. Compared to previous RNA-seq pipeline evaluation studies that were restricting on genetic variations (e.g., gene fusions) [131], we have extended the scope to downstream applications such as survival predictions using gene expression. The results revealed that RNA-seq pipelines jointly impact the gene expression estimation, and the influence will carry on to downstream applications. Most importantly, we found that RNA-seq pipelines that produced more accurate gene expression resulted in better survival prediction performance.

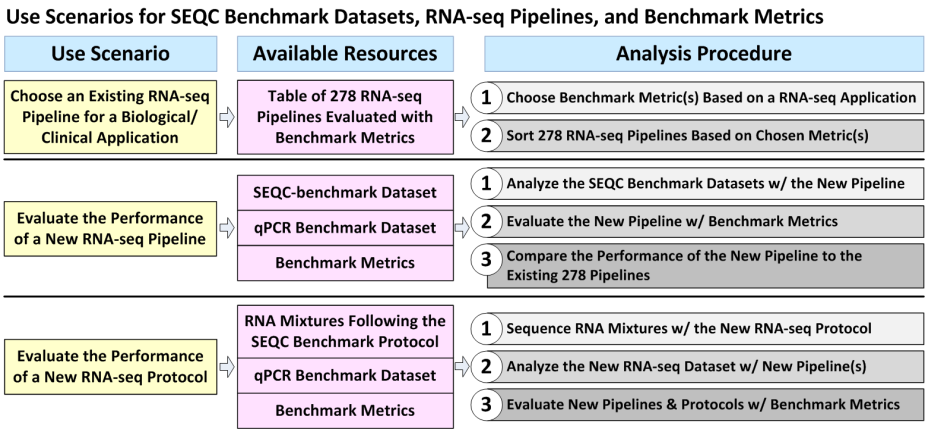
Putting our results in the context of real-world application, we offer the following scenarios by taking advantage of our findings:

In scenario 1 where researchers need to select a pipeline to analyze an Illumina dataset (or similar short-read sequence dataset), they may refer to our evaluation results to choose a pipeline by following these steps (Figure 3.6): (1) Select a metric based on the requirements of the clinical application (not necessarily predictive modeling). (2) Sort the pipelines based on this metric and choose the top pipeline. For example, depending on the application, different weights can be assigned to the three metrics instead of using the average rank. Researchers who want to conduct initial filtering of genes to identify DEGs may want to stress the importance of correct quantification of relative gene expression. Thus, they may want to focus on the accuracy metric. Top and bottom pipelines in terms of accuracy are listed in Figure 3.6a. Median normalization is the frequently occurring component in the

Good-Performing RNA-seq Pipelines for Various Applications

RNA-seq Application	Metric	RNA-seq Pipelines
Accurate estimation of relative gene expression with benefit for differentially expressed gene detection	**Accuracy** (Deviation from qPCR)	• Bowtie + RSEM + Median • Bowtie2 Single-Hit + [Count-Based/Cufflinks/RSEM] + Median • Bowtie2 Multi-Hit + RSEM + Median • BWA + [Count-Based/RSEM] + Median • GSNAP Spliced [Single-/Multi-Hit] + [Count-Based/Cufflinks] + Median • GSNAP Un-spliced Multi-Hit + RSEM + Median • MAGIC [Single-/Multi-Hit] + Count-Based + Median • OSA + [Count-Based/Cufflinks] + Median • STAR + Count-Based + Median • TopHat [Single-/Multi-Hit] + [Count-Based/Cufflinks] + Median • WHAM Single-Hit + Count-Based + Median • WHAM Multi-Hit + RSEM + Median
Small variation in gene expression across all replicate libraries for a single sample	**Precision** (Coefficient of variation across replicate libraries)	• Bowtie2 Multi-Hit + Count-Based + [FPM/FPKM/Upper Quartile] • Bowtie2 Multi-Hit + Cufflinks + RLE • GSNAP Un-Spliced [Single-/Multi-Hit] + [Count-Based/Cufflinks] + [FPM/FPKM/Upper Quartile/RLE] • GSNAP Un-Spliced [Single-/Multi-Hit] + Cufflinks + TMM
Small within-sample variation in gene expression across all replicate libraries compared with between-sample variation	**Reliability** (Intraclass [intra-sample] correlation for grouped data)	• Bowtie2 [Single-/Multi-Hit] + [Count-Based/Cufflinks/RSEM] + Median • BWA + [Count-Based/Cufflinks/RSEM] + Median • GSNAP Spliced [Single-/Multi-Hit] + [Count-Based/Cufflinks] + Median • MAGIC [Single-/Multi-Hit] + Count-Based + Median • MapSplice + [Count-Based/Cufflinks] + Median • Novoalign [Single-/Multi-Hit] + [Count-Based/Cufflinks] + Median • OSA + Cufflinks + Median • RUM + Count-Based + Median • Subread + Count-Based + Median • TopHat [Single-/Multi-Hit] + [Count-Based/Cufflinks] + Median

(a) Good-performing RNA-seq pipelines for various applications.



(b) Use scenarios for SEQC benchmark datasets, RNA-seq pipelines, and benchmark metrics.

Figure 3.6: The resources provided by this study (i.e., the 278 RNA-seq pipelines, the benchmark metrics, and the SEQC-benchmark datasets) can serve as guidelines for biological and clinical researchers as well as for bioinformaticians and biotechnologists. (a) Depending on the gene expression application, the three metrics (i.e., accuracy, precision, and reliability) may be used to choose a pipeline. We have associated each metric with an RNA-seq application and listed the top-performing pipelines for each metric. The red-highlighted component in each listed RNA-seq pipeline indicates components that frequently occur among the top-performing pipelines for each metric. (b) Biological or clinical researchers who want to analyze Illumina RNA-seq data (or data from similar platforms with short, fixed-length reads) can choose an existing RNA-seq pipeline using the provided table of 278 pipelines ranked by accuracy, precision, or reliability. Bioinformaticians that are developing a new RNA-seq pipeline for Illumina data (or data from similar platforms) can use the SEQC-benchmark datasets and benchmark metrics to evaluate the new pipeline and assess its performance relative to the 278 pipelines. Bioinformaticians or biotechnologists that are developing new RNA-seq protocols can first sequence the same RNA mixture samples (i.e., samples A, B, C, and D), and then evaluate associated data analysis pipelines using the qPCR benchmark dataset and the benchmark metrics.

most accurate pipelines. If experimental conditions limit the sample size, small variation among replicate libraries may be important to better estimate gene expression. The precision and reliability metrics were designed to capture variation in gene expression from different perspectives. Researchers who need small variation in gene expression across replicate libraries for a single sample may want to weight the precision metric. Bowtie2 multi-hit and GSNAP un-splice mapping are the frequently occurring components in the most precise pipelines. Researchers who need small within-sample variation in gene expression relative to between-sample variation may want to emphasize the reliability metric. Again, median normalization is the frequently occurring component in the most reliable pipelines.

In scenario 2 where researchers want to evaluate a newly developed RNA-seq pipeline for an Illumina dataset, they may use the SEQC-benchmark dataset, qPCR benchmark dataset, and the three metrics as follows (Figure 3.6b): (1) Analyze the SEQC-benchmark dataset with the new RNA-seq pipeline. (2) Evaluate the new pipeline with the three metrics. (3) Compare the performance of the new pipeline to the 278 RNA-seq pipelines. The 278 RNA-seq pipelines serve as a representative sample, or benchmark, of pipelines for gene expression estimation that include both good-performing and poor-performing pipelines.

In the third scenario, researchers who want to evaluate a new RNA-seq protocol may use the RNA samples from the SEQC protocol (i.e., samples A, B, C, and D), the qPCR benchmark dataset, and the three metrics as follows: (1) Sequence the RNA samples with the new RNA-seq protocol. (2) Analyze the new dataset with new pipelines specifically designed for the new RNA-seq protocol. (3) Evaluate the new pipelines and the new RNA-seq protocol with the three metrics. Ideally, existing pipelines could be applied as fixed variables to both old and new RNA-seq protocols to attribute any changes in gene expression estimation performance to the change in RNA-seq protocols.

It is worthwhile to mention that, in the Phase-II experiment, we set a threshold to divide

patients in each dataset into two groups to conduct downstream predictions. This choice balances the numbers of samples in two groups and balances the computational complexity with the biological meaningfulness. With RNA-seq pipeline evaluation as the main focus of this study, we want to minimize the sample size imbalance introduced bias in the down-stream biological applications. The investigation can be improved if the choice of threshold is more towards the real-world clinical scenarios. Moreover, we did not investigate the underlying genes that contribute to the prediction. The biological and medical contexts of the gene signature are critical for the causality assessment and clinical application in explaining the disease outcome predictions. Thus, more sophisticated down-stream applications of RNA-seq can be explored in the future to evaluate the impact of RNA-seq pipelines. In summary, we showed that upstream RNA-seq pipelines that performed well in gene expression estimation generally performed well for downstream gene expression-based prediction of disease outcome. Our study represents a large-scale and objective assessment of the predictive performance of various RNA-seq pipelines and should prove useful in moving RNA-seq-based predictive models closer to clinical applications.

3.2 Multi-Omics Integration with Cross-Modality Translation

3.2.1 Background

Breast cancer is the most common type of cancer in females worldwide. In 2018, breast cancer constituted over 25% of about 8.5 million new cancer diagnoses in female patients [167]. This prevalence pattern is found in the US as well, where women have over a 12% risk of being diagnosed with breast cancer in their lives, and breast cancer cases are expected to encompass about 30% of new cancer cases. While the principal risk factor for breast cancer is age, it is known that selected gene mutations account for about 10% of all breast cancer cases. Research into prognostic genomic biomarkers beyond mutational status is ongoing and may offer insights into disease mechanisms and new therapies. Breast cancer maintains the second-highest mortality rate for cancers in females at about 13%.

Survival rates for breast cancer are typically measured by 5-year post-diagnosis survival. The 5-year survival rate is 90% when all stage classifications are considered. With stage breakdown accounted for, the risk can be further stratified, as localized breast cancer survival rate is 99%, while this drops to 85% and 27% for regionally and distantly spread cancer, respectively.

Machine learning for cancer survival analysis has attracted increasing attention in recent years. Public multi-omics datasets such as The Cancer Genome Atlas (TCGA) [168] have greatly accelerated the research for survival analysis using -omics data. The survival analysis can be categorized into binary classification or risk regression. In a binary classification task, the patients are typically split into a short-survival group and a long-survival group based on a predefined threshold (e.g., five years). While in risk regression studies, each patient's survival time is also taken into account, which is usually modeled with the Cox proportional hazards model [169] and its extensions.

Various methods have been employed with great success in developing survival prediction models with large and heterogeneous cancer datasets. For example, Zhao et al. tested various classification algorithms to predict 5-year breast cancer survival by integrating gene expression data with other clinical and pathological factors [170]. Authors found that all methods tested, including gradient boosting, random forest, artificial neural networks, and support vector machine, performed rather similarly with accuracy and area under the curve (AUC) of $.72$ and $.67$, respectively. Importantly, this study demonstrates that classification methods may not matter as much as the quality of the data itself [170]. Goli et al. developed a breast cancer survival prediction model with clinical and pathological data using support vector regression and found similar positive results [171]. This study establishes the use of support vectors as a promising route in survival prediction with imbalanced datasets. Similarly, Gevaert et al. integrated microarray gene expression data with clinical data using Bayesian Networks and achieved a maximum AUC of $.845$ [76]. Importantly, this study found that incorporating both data modalities improved predictions beyond either clinical

or gene expression alone. Sun et al. created prediction models for 5-year breast cancer survival using genomic data, including gene expression, copy number alteration, methylation, and protein expression, coupled with pathological imaging data from TCGA. The authors utilized multiple kernel learning to enact feature-level integration of all data. Their multi-omics model, excluding imaging data, had an AUC of 0.802 ± 0.032. When incorporating the imaging data, the AUC went up slightly to 0.828 ± 0.034 [172]. Ma et al. have applied factorization autoencoder to integrate gene expression, miRNA expression, DNA methylation, and protein expression for progression-free interval event prediction and achieve an AUC of 0.74 on bladder cancer and an AUC of 0.825 on brain glioma [81].

Instead of binary classification, the survival risk regression aims to predict the expected duration of time until one or more events happen by modeling the time to event data. The proportional hazards model assumes the covariates are multiplicatively related to the hazard [173]. Assuming the proportional hazards assumption holds, the Cox proportional hazards model can estimate the effect parameters without any consideration of the hazard function [169]. With the development of deep learning, the Cox proportional hazards model has been extended with deep neural networks. For example, Deep Surv [174] and Cox-Time [175] replace the linear relationship in the Cox proportional hazards model with non-linear neural networks. In addition, L_1 and L_2 regularization terms have been utilized on the network parameters to reduce the over-fitting of the models. The survival regression model has also been applied to multi-omics data. For example, Huang et al. have developed a Cox-proportional hazards model based multi-omics neural network for breast cancer survival regression [82].

In our previous study [83], we have built a transnational pipeline for overall survival prediction of breast cancer patients by decision-level integration of multi-omics data (e.g., gene expression, DNA methylation, miRNA expression, and copy number variations (CNVs)). However, many right-censored samples have been discarded to enable binary classification. In this study, we extended the work by replacing the binary survival classification with sur-

vival risk regression to make the most of the TCGA dataset [176]. We hypothesize there are both complementary and consensus information in the multi-omics data. To utilize the complementary and consensus information among multi-omics data, we replace the decision-level integration with deep learning-based feature-level integration [176]. The remainder of the section is structured as follows: in subsection 2, we first describe the simulated two-view data from the Modified National Institute of Standards and Technology (MNIST) database and multi-omics breast cancer (BRCA) data from the TCGA database (referred as TCGA-BRCA hereafter). We then present the proposed methods for multi-omics data integration by utilizing the complementary information and consensus information among modalities. In subsection 3, we present the results of the baseline models and proposed models on both MNIST simulated data and TCGA-BRCA multi-omics data. We will discuss the results and conclude the work in subsection 4 and subsection 5, respectively.

3.2.2 Methods

Simulated Multi-View MNIST Dataset

We simulate the multi-modality data from the Modified National Institute of Standards and Technology (MNIST) database to validate the proposed feature-level integration network. The MNIST database consists of 60,000 training samples and 10,000 testing samples. Each sample in the MNIST database is a 28×28 image of a single hand-written digit from 0 to 9. The goal is to train a multi-class classifier to predict the digit from the input image.

We simulate two-views of each hand-written digit image from the MNIST database (Fig. 3.7A). The first view (X_1) is the original image from the MNIST database, while the second view (X_2) is the corresponding rotated image (90-degree counter-clockwise rotation). We further simulate noises for the data because the task is easy even for single-view data. We have simulated two kinds of noises and apply them to both views of the hand-written digit images: pixel-wise Gaussian noise (Fig. 3.7B) and random erasing (Fig. 3.7C).

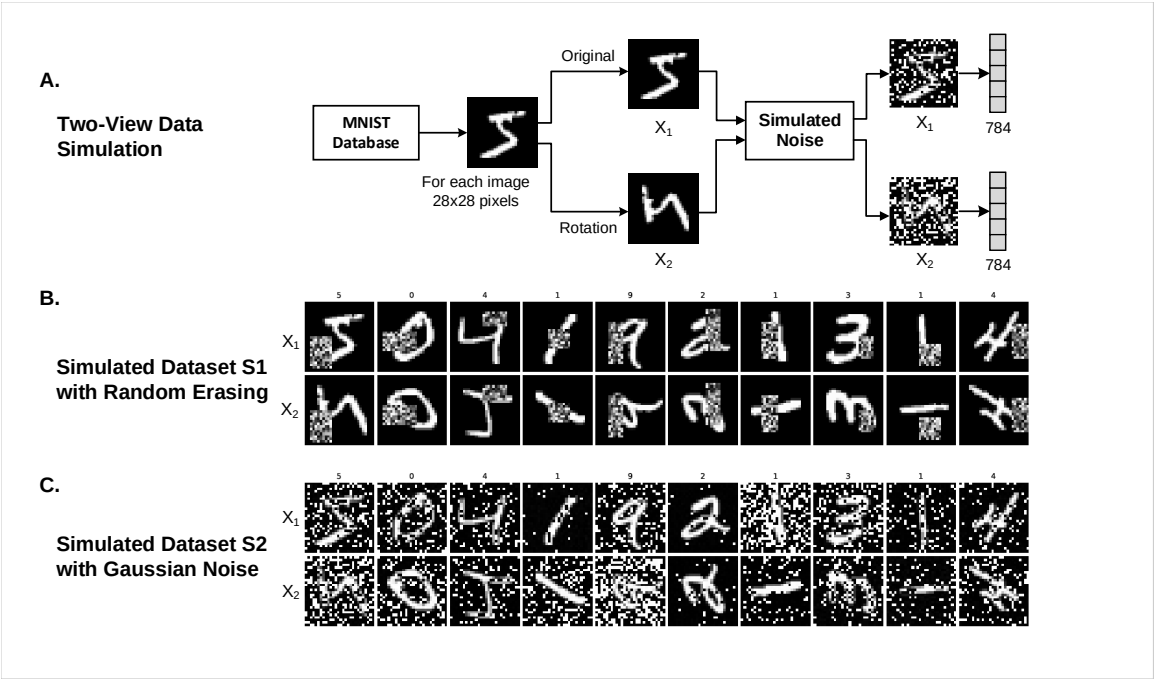


Figure 3.7: Simulation two-view data from MNIST database. A. Pipeline for simulation of two-view data from the MNIST database. B. Simulated dataset S_1 with random erasing noise. C. Simulated dataset S_2 with Gaussian noise.

We flatten the image to a vector with a length of 784 as the final input to deep neural networks.

TCGA-BRCA Breast Cancer Multi-Omics Dataset

We obtain breast cancer data from the TCGA database [168], which is a public database containing genomic data for over 20,000 paired cancer and normal samples from 33 cancer types. By keeping the samples that are simultaneously profiled with gene expression, miRNA expression, DNA methylation, and CNVs, the final dataset size for survival analysis is 1,060.

We summarized the four omics data modalities obtained from the TCGA-BRCA dataset in Table 3.1 For gene expression, the number of features includes different isoforms for each gene and some non-coding RNA transcripts. The DNA methylation beta value ranges from 0 to 1, where a beta value of 0 means that no methylation is detected for that probe,

Table 3.1: Overview of four omics data modalities

Data Modality	Gene Expression	DNA Methylation	miRNA Expression	Copy Number Variation
Measures	Fragments per killobase of transcript per million mapped reads (FPKM)	Beta Value	Reads per million mapped reads (RPM)	Gain/Loss/Neutral
Dynamic Range	Continuous [0, 3823803664.0]	Continuous [0, 1]	Continuous [0, 679286.5]	Discrete {"Loss": -1, "Neutral": 0, "Gain":1}
Feature Name	Ensembl Gene ID	cg probe identifier	miRNA identifier	Ensembl Gene ID
# of Features	60,483	25,978	1,881	19,729

while a 1 means that the CpG was always methylated. For CNV features, "Gain" means more copies of a gene than normal, while "Loss" means fewer copies of a gene than normal.

The overall pipeline for multi-omics survival analysis is presented in Fig. 3.8. Quality control and preprocessing are essential for making sense of multi-omics data. To get rid of the low-quality features, we remove features with missing data. For the gene expression and miRNA expression data, we also apply a log transform $\log_2(X + 1)$ to the features, where X is the FPKM for gene expression and RPM for miRNA expression. We then apply min-max normalization to scale all four data modalities to a range of 0 to 1. After the quality control and normalization, we apply a stratified four-fold split of the data into a training set (60%), validation set (15%), and a testing set (25%) in each fold.

The multi-omics data usually suffer from the "curse of dimensionality," where the number of features is significantly larger than the number of samples. To mitigate this challenge, researchers usually apply feature selection or dimension reduction techniques to get rid of the unrelated or redundant features, which are essential for the success of downstream analysis such as classification or survival analysis. For classification, supervised univariate feature selection methods such as minimum Redundancy Maximum Relevance (mRMR) [177] and mutual information can be used. For survival analysis, it is not straightforward to apply supervised feature selection, and thus various unsupervised or knowledge-guided feature selection has been applied. For example, Huang et al. have applied gene

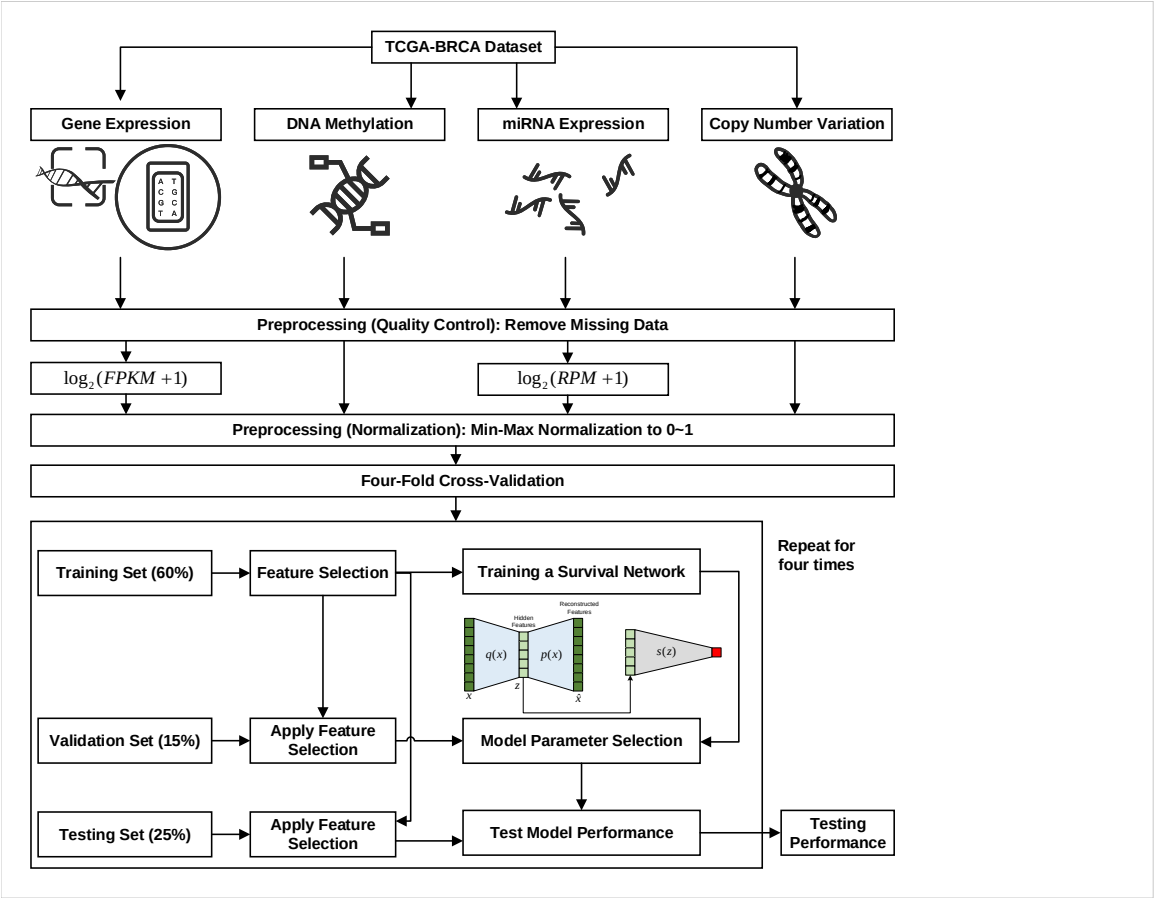


Figure 3.8: Overall pipeline for survival analysis. We obtain multi-omics data (i.e., gene expression, DNA methylation, miRNA expression, and copy number variation) for breast cancer patients from the TCGA-BRCA database. The multi-omics data are preprocessed and normalized to a range of 0 to 1. We then apply four-fold cross-validation and split the data into a training set (60%), validation set (15%), and testing set (25%) in each fold. We train the feature selection or dimension reduction step and the survival networks using the training set and apply them to the validation set for parameter selection and the testing set for performance reporting.

co-expression analysis as the dimension reduction approach [82]. This study focuses on deep-learning-based feature-level integration. Thus, we have applied two simple feature selection/dimension reduction approaches, principal component analysis (PCA) and unsupervised variance-based feature selection. We apply PCA to the training dataset for PCA-based dimension reduction and use the first 100 principal components (PCs) of training, validation, and testing datasets for survival analysis. We select the top 1000 features from the training dataset with the highest variances for unsupervised variance-based feature selection. We then use these 1000 features of training, validation, and testing datasets for survival analysis.

Single-Modality Network

For single-modality data, we propose to use an autoencoder and a task-specific network for single-modality classification or survival analysis (Fig. 3.9). For the input data x after feature selection, we first apply an encoder $q(x)$ to transform the input data to a hidden feature z, and then reconstruct the input data $\hat{x}$ from the hidden feature with a decoder $p(z)$. We then feed the hidden feature z into a task-specific network for classification or survival analysis.

Endpoint 1: Multi-Class Classification For the classification network $c(z)$, we use a fully connected network with the output dimension the same as the number of classes. Thus, the whole network is trained with reconstruction loss L_{recon} and the classification loss L_{cls}. In this study, we use the mean-square error for the reconstruction loss:

$$L_{recon} = \frac{1}{N}\sum_{1}^{N}(x_n - \hat{x}_n)^2 \tag{3.1}$$

where N is the batch size. We use the cross-entropy loss for the classification loss:

$$L_{clf} = -\log(\frac{\exp(x[class])}{\sum_{j=1}^{C}(x[j])}) = -x[class] + \log(\sum_{j=1}^{C}(x[j])) \tag{3.2}$$

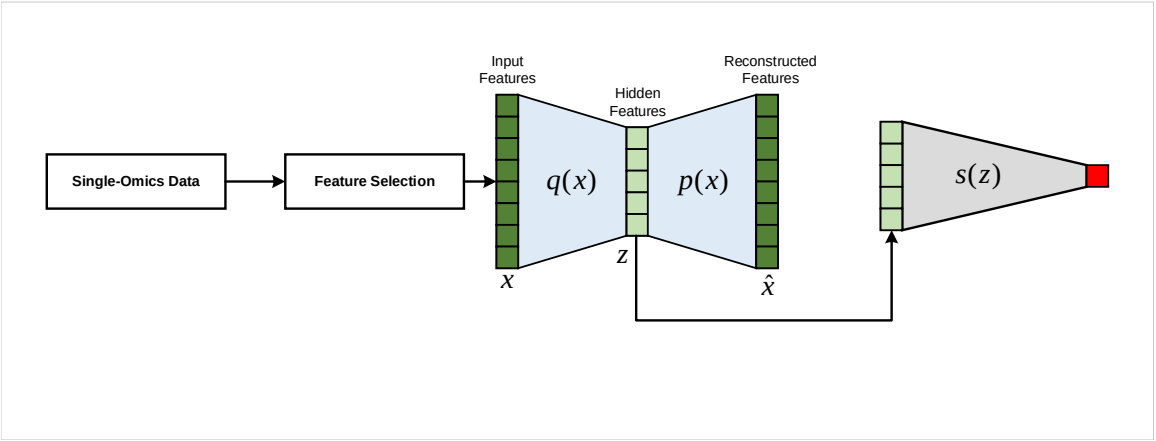


Figure 3.9: Single-omics data survival analysis network. The input data x is represented with an encoder $q(x)$ into hidden feature z and then constructed with a decoder $p(x)$. We then feed the hidden feature z into a task-specific network such as multi-class classification or survival analysis.

where C is the number of classes and $j \in 1, ..., C$. For each epoch, we first train the encoder-decoder with the reconstruction loss L_{recon} and then train the encoder and classification network with the cross-entropy loss L_{clf}.

The multi-class classification performance is evaluated by accuracy, weighted precision, and weighted recall. These metrics are in the range of $[0, 1]$, and the higher the better. We do not include AUC as a metric because we perform 10-class classification with the simulated MNIST dataset instead of binary classification.

Endpoint 2: Survival Analysis We also use a fully connected neural network $s(z)$ to replace the Cox proportional hazards model for the survival analysis. The output of the survival network $s(z)$ is the patient's hazard h. Based on the Cox proportional hazards model, the survival network is trained with the negative log partial likelihood loss L_{sur}:

$$L_{sur} = -\frac{1}{N_{ob}} \sum_{i:C_i=1} (h_i - \log \sum_{j:T_j \geq T_i} \exp(h_j)) \tag{3.3}$$

where $C_i = 1$ indicates the occurrence of the event for patient i, N_{ob} is the total number of events in the batch, and T_i and T_j are the survival time for patient i and patient j, respectively.

For the evaluation of the risk scores predicted by survival models, various metrics have been developed to measure the concordance between the predicted risk scores and the actual survival time. Following the previous studies in deep-learning-based survival analysis [82], we will evaluate the overall survival analysis performance with the concordance index (C-index) [178]. C-index evaluates how well the survival risk we computed aligns with the actual survival time given any two comparable pairs:

$$\text{C-index} = \Pr\{h_i > h_j | T_i < T_j, C_i = 1\} \tag{3.4}$$

Multi-Modality Integration Network

There are two principles in multi-view learning: 1) the complementary principle assumes that each view contains information other views do not have, and we should extract the difference from each view while preserving the common information; 2) the consensus principle assumes that the disagreement between views upper bounds the classification errors; thus, we should aim to maximize the agreement between views. Based on these two principles, we have the following approaches for integrating data from multiple -omics modalities.

1) Integrating the Complementary information: Concatenation Autoencoder (ConcatAE). Similar to the methods discussed in Chapter 2, we use the concatenation autoencoder (ConcatAE) to integrate the complementary information from each data modality (Fig. 3.10). For each modality, we train an independent autoencoder and transform the input features into a hidden space. We then concatenate the hidden features from each modality and feed the concatenated hidden feature into the task-specific network. Compared to the single-modality network, we have a separate reconstruction loss for each data modality. Thus, the reconstruction loss is the summation of these separate reconstruction losses. For example,

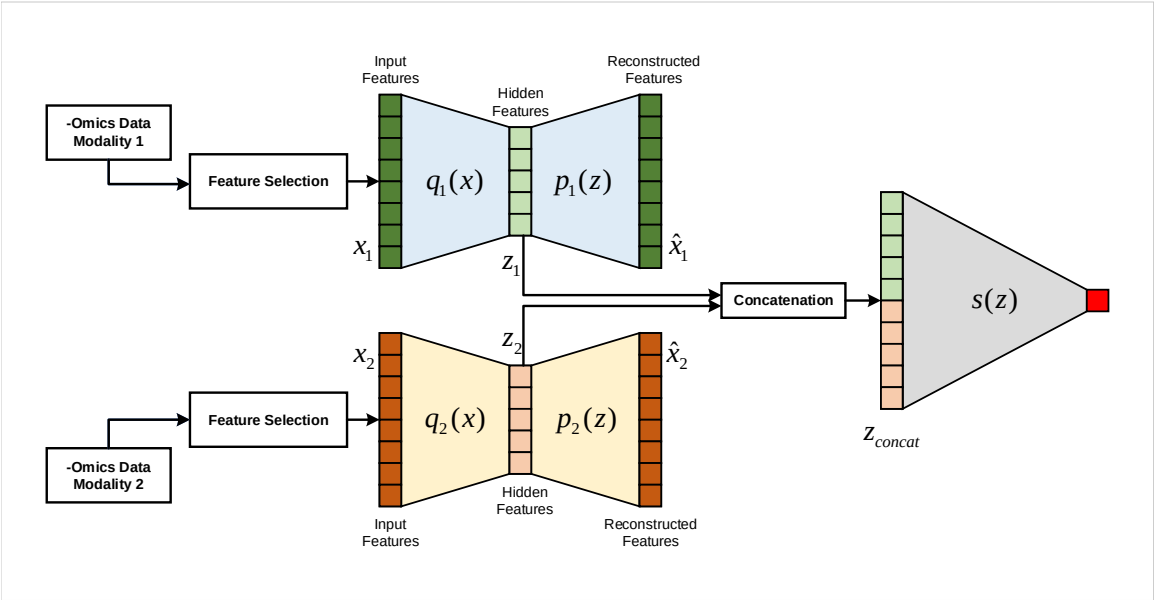


Figure 3.10: Multi-omics data integration with concatenation autoencoder (ConcatAE). The hidden features of each data modality are concatenated before feeding into the task-specific network.

when integrating two modalities, the new reconstruction loss would be:

$$L'_{recon} = \frac{1}{N}\sum_{1}^{N}((x_{1,n} - \hat{x}_{1,n})^2 + (x_{2,n} - \hat{x}_{2,n})^2) \tag{3.5}$$

The task-specific network training procedure remains the same, with the input becoming the concatenation of hidden features represented from each modality.

2) Integrating the Consensus Information: Cross-Modality Autoencoder (CrossAE). We use the cross-modality autoencoder (CrossAE) to integrate the consensus information from each data modality (Fig. 3.11). The key idea to enable consensus representation among modalities is using the hidden features represented from one modality to reconstruct the input features from other modalities.

We train the framework with three steps. In the first step, we train an autoencoder for each modality independently, as we have done in the ConcatAE model with L'_{recon}. In the second step, we train these encoders and decoders again with cross-modality reconstruction. For example, the modality 1 encoder encoder $q_1(x)$ is used to transform input data x_1

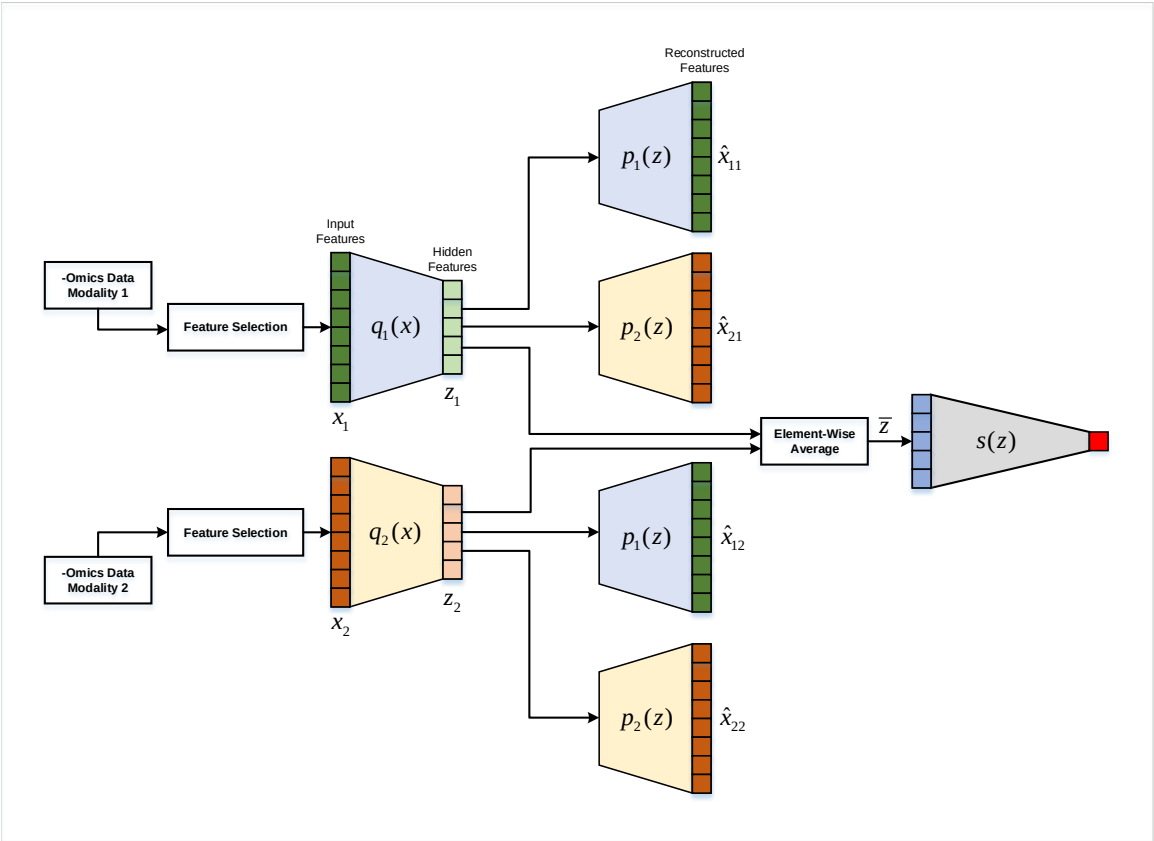


Figure 3.11: Multi-omics data integration with cross-modality autoencoder (CrossAE). For hidden features of each data modality, they are used to reconstruct input features of both the original modality and other modalities. The hidden features of various modalities are element-wise averaged before feeding into the task-specific network.

to hidden feature $z_1 = q_1(x_1)$. We then use the modality 2 decoder $p_2(z)$ to reconstruct the modality 2 input data x_2 from z_1, which is denoted as $\hat{x}_{21} = p_2(z_1)$. We can perform similar cross-modality reconstruction from modality 2 hidden features z_2 to modality 1 input data x_1. Thus, the cross-modality reconstruction loss L_{cross_recon} for step 2 with two modalities is

$$L_{cross\_recon} = \frac{1}{N}\sum_{1}^{N}((x_{1,n} - \hat{x}_{12,n})^2 + (x_{2,n} - \hat{x}_{21,n})^2) \tag{3.6}$$

In the third step, we combine the hidden features from each modality with the element-wise average and then train the encoders and task-specific network with task-specific loss (e.g., the cross-entropy loss for classification or the negative partial log-likelihood loss for survival regression).

We only implemented and tested the proposed integration models on two data modalities, but we believe these frameworks can be naturally extended to the integration of more than two data modalities.

Implementation and Experiments

The train-test split for cross-validation and the classification metrics are implemented with scikit-learn [179]. The neural networks are designed and implemented with PyTorch 1.1.0. For cancer type classification, we use a batch size of 32, an Adam optimizer with a learning rate of 0.001, and training epochs of 200. For survival analysis, we use a batch size of 128, an Adam optimizer with a learning rate of 0.001, and training epochs of 200. More details of the model implementation and training details can be found at Github repo (https://github.com/tongli1210/BreastCancerSurvivalIntegration).

Table 3.2: Multi-modality integration simulation with MNIST dataset

	Random Erasing (S_1)			**Gaussian Noise (S_2)**		
Modalities	**ACC**	**Precision**	**Recall**	**ACC**	**Precision**	**Recall**
X1	0.942±0.004	0.942±0.004	0.942±0.004	0.884±0.003	0.886±0.003	0.884±0.003
X2	0.942±0.003	0.943±0.003	0.942±0.003	0.879±0.005	0.881±0.005	0.879±0.005
ConcatAE(X1+X2)	**0.962±0.001**	**0.963±0.001**	**0.962±0.001**	0.924±0.001	0.925±0.002	0.924±0.001
CrossAE(X1+X2)	0.962±0.002	0.962±0.002	0.962±0.002	**0.933±0.002**	**0.933±0.002**	**0.933±0.002**

3.2.3 Results

Multi-Modality Integration Simulation

We first test the proposed single and multi-modal integration networks on the simulated MNIST datasets (S_1 and S_2). The results are presented in Table 3.2. From the results, we can observe significant classification performance improvements after multi-modality data integration for both random erasing dataset S_1 and the Gaussian noise erasing dataset S_2. For dataset S_1, we assume the model should take the complementary information from X_1 and X_2 to get better performance. From the experiment results, the integration model ConcatAE performs slightly better than the integration model CrossAE. For dataset S_2, because of the global noises for both views, we assume the model should take the consensus information from S_1 and S_2 to get better performance. From the experiment results, we observe CrossAE achieves better performance compared to ConcatAE, which is as expected.

Multi-Modality Integration for Breast Cancer Survival Analysis

The performance of the single-omics survival analysis model is presented in Table **??**. We observe that the model achieves better performance when using PCA features than using the high variance features for all modalities except for CNVs. Among the four -omics data, miRNA expression is the most predictive for overall survival, followed by DNA methylation and gene expression. Moreover, CNVs are the least predictive for breast cancer overall survival, which is consistent with our previous findings [83]. The best single-omics survival analysis performance is a C-index of 0.616 ± 0.057, achieved by miRNA

Table 3.3: Performance of single-omics survival analysis model

Data Modality	Gene Expression	DNA Methylation	miRNA Expression	Copy Number Variation
PCA	0.589±0.084	0.583±0.058	**0.616±0.057**	0.476±0.051
Variance	0.529±0.033	0.581±0.066	**0.614±0.041**	0.503±0.071

Table 3.4: Performance of multi-omics survival analysis model

Integration	Data Modality	GeneExp + DnaMeth	GeneExp + miRNA	GeneExp + CNVs	DnaMeth + miRNA	DnaMeth + CNVs	miRNA + CNVs
ConcatAE	**PCA**	0.585±0.107	0.59±0.093	0.576±0.047	**0.641±0.031**	0.583±0.09	0.588±0.057
	Variance	0.507±0.036	0.53±0.052	0.524±0.038	0.625±0.023	0.586±0.068	0.603±0.04
CrossAE	**PCA**	0.583±0.07	0.595±0.062	0.553±0.045	0.63±0.081	0.579±0.065	0.578±0.028
	Variance	0.511±0.027	0.558±0.054	0.53±0.033	0.605±0.059	0.576±0.026	0.613±0.066

data with PCA features.

The performance of the multi-omics integration survival analysis model is presented in Table 3.4. Based on the results, we can observe that integration is not always beneficial for performance. For example, the integration of gene expression and DNA methylation high variance features can lead to a lower C-index (0.507 ± 0.036) than either gene expression (0.529 ± 0.033) or DNA methylation (0.581 ± 0.066) alone. Among the six combinations of two-omics data integration, we found the integration of DNA methylation and miRNA expression consistently achieves a good performance. Comparing the two integration strategies, we found that the ConcatAE outperforms the CrossAE in most experiments. Comparing the two feature selection strategies, we observed that the PCA features outperform high variance features in most experiments except for those involves CNV data. We believe the PCA dimension reduction approach may not be suitable for the discrete CNV data. Among all multi-omics integration models, the best performance (0.641 ± 0.031) is achieved by integrating DNA methylation and miRNA expression using PCA features and the ConcatAE model.

To evaluate the consensus among hidden features, we measure the similarity of paired hidden features with the Euclidean distance and visualize their distributions with grouped violin plots in Fig. 3.12. The violin plots are grouped by multi-omics modalities under

integration (e.g., GeneExp+miRNA) and compared for the two integration methods ConcatAE and CrossAE. For the hidden features (dimension of 10) represented from PCA features, we can observe higher similarities (or lower Euclidean distances) for integration using CrossAE compared to those using ConcatAE (Fig. 3.12A). However, for the hidden features (dimension of 100) represented from high variance features, the CrossAE method will not necessarily lead to higher similarities (Fig. 3.12B). The observation is further confirmed with grouped bar plots of the average Euclidean distances in Fig. 3.12C and Fig. 3.12D. The results indicate that the consensus constraints imposed by CrossAE work well for PCA features but suffer for the high variance features, which have a much higher dimension.

To further understand the similarity among paired hidden features, we have also visualized the hidden features from the first fold of our four-fold cross-validation with the t-Distributed Stochastic Neighbor Embedding (t-SNE). Based on the t-SNE visualization, for hidden features represented from PCA features, we can observe better overlaps of the CrossAE features (Green and Yellow) compared to those of the ConcatAE features (Red and Blue), which indicate the effect of consensus constraints on multi-omics data representation. However, we observe similar patterns for the ConcatAE features (Red and Blue) and the CrossAE features for hidden features represented from high variance features (Green and Yellow). Thus, for the high variance features, the effect of consensus constraints by CrossAE is not as significant.

3.2.4 Discussions

In this study, we have developed two multi-modal data integration strategies. We propose to integrate the complementary information among modalities with ConcatAE and integrate the consensus information using CrossAE. We have tested the proposed models on the simulated MNIST data and validated their effectiveness. We then apply the proposed models to the multi-omics breast cancer survival data. ConcatAE model integrating DNA methyla-

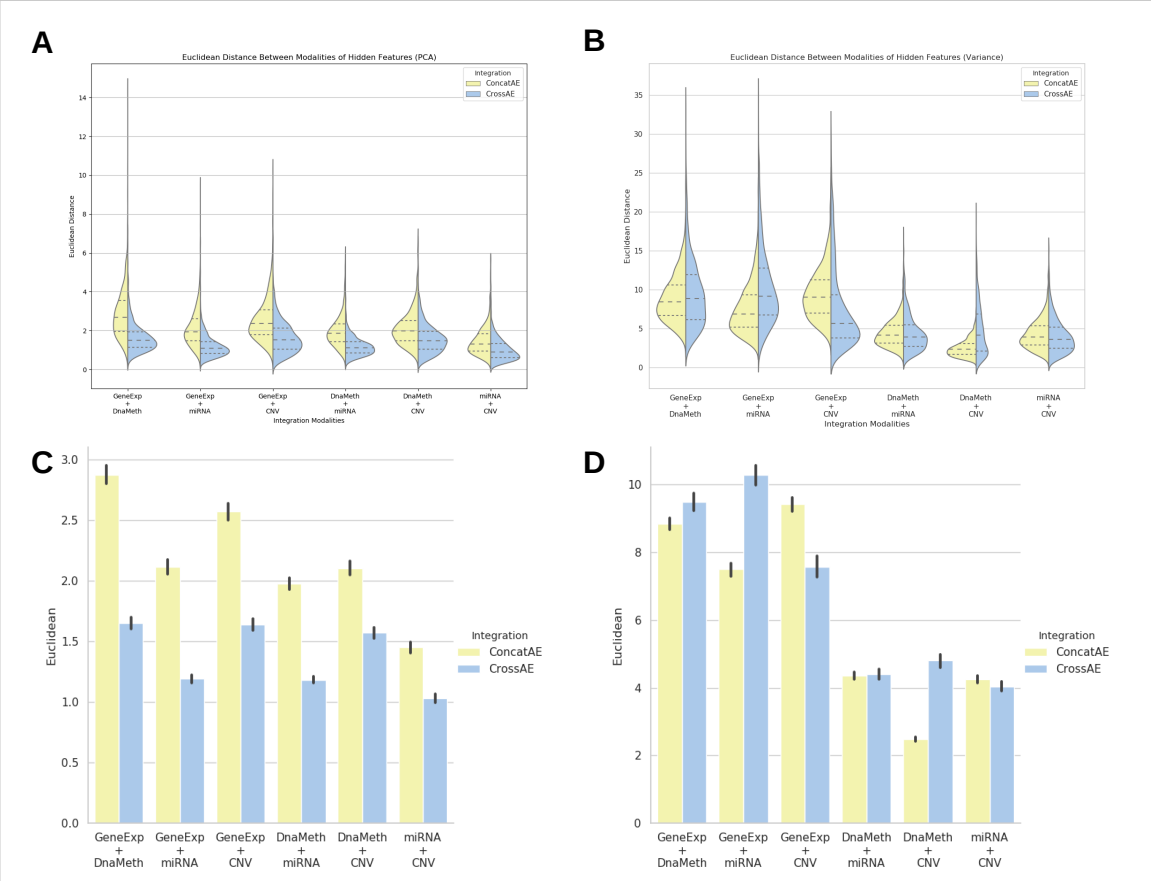


Figure 3.12: Similarity measure with Euclidean distance of the paired hidden features. We measure the similarity of paired hidden features with the Euclidean distance. A. Grouped violin plots of the Euclidean distances for hidden features represented from PCA features. B. Grouped violin plots of the Euclidean distances for hidden features represented from high variance features. C. Grouped bar plots of the average Euclidean distances for hidden features represented form PCA features. D. Grouped bar plots of the average Euclidean distances for hidden features represented from high variance features. Yellow: ConcatAE. Blue: CrossAE.

tion and miRNA expression PCA features achieves the best performance with a C-index of 0.641 ± 0.031 and outperforms that of the CrossAE model (0.63 ± 0.081). Both integration approaches outperform the corresponding single-modality model, which uses DNA methylation or miRNA expression alone. The results indicate that these two modalities should have both complementary and consensus information for survival prediction.

Although the ConcatAE outperforms CrossAE, we believe this does not necessarily indicate that the complementary information is more important than the consensus information. As we have seen in the MNIST simulated data with Gaussian noise, if the multi-modality data are noisy and equally predictive, consensus learning can achieve higher prediction performance than that of complementary learning. Moreover, the ConcatAE model should include both the modality-invariant and modality-unique information, although neither has explicitly been maximized.

The best survival prediction performance is achieved by the integration of DNA methylation and miRNA expression PCA features. However, the results are insufficient to conclude that DNA methylation or miRNA expression is more informative than the other modalities. Because of the lack of biological ground truth, the model interpretation and wet-lab validation are needed to understand the model. As a black-box model, we cannot currently locate which biomarkers (e.g., specific genes or methylation sites) are picked by the integration network and contribute most to the final survival prediction. Thus, as a future direction, we propose to apply model interpretation methods to the integration networks and identify the biomarkers picked by the deep network. Then, these biomarkers can be validated by literature if already discovered or by wet-lab experiments if novel. The identification and validation of biomarkers can provide direct evidence on why some integration models outperform the others, which is essential to better understand the results and allow translational impact on the clinical applications.

The TCGA-BRCA dataset we used for multi-omics integration is another major factor that might influence the survival prediction performance. Based on our results, the

CNV features are the least predictive for breast cancer survival analysis. However, the CNV features we obtained from the TCGA database are categorical (i.e., "gain", "loss", or "normal"). The simplified CNV features might constrain the predictive capability of this modality. Moreover, the gene expression data we downloaded from TCGA are normalized with FPKM, and the miRNA expression data are normalized with RPM. However, the FPKM and RPM normalization are potentially biased when comparing between samples. Although the feature extraction with RNA-seq pipelines is out of this study's scope, we believe the survival prediction performance can be further improved for gene expression and miRNA expression if replacing the normalization method with more sophisticated techniques such as transcripts per million (TPM).

Although we have demonstrated the effectiveness of ConcatAE and CrossAE for multi-omics data integration, the current study has several limitations and can be improved in the follow-up studies. We will discuss these future directions in the model validation dataset, generalization to larger datasets with various endpoints, and model improvements.

This study validates the proposed ConcatAE and CrossAE networks' effectiveness with the simulated two-view imaging data from the MNIST database. We have controlled and visualized the consensus and complementary information. Ideally, a cancer genomics dataset with ground truth would be preferred to validate the proposed integration networks. However, based on our knowledge, there is no such golden standard multi-omics dataset developed yet. As we do not fully understand the complex interactions among multi-omics data, it is currently infeasible to annotate all of them and develop a real-world dataset with ground truth. We believe it is worth the biological communities to develop golden standard multi-omics datasets to mitigate this challenge. It will be helpful to validate the consensus and complementary principles for computational multi-omics data integration methods. For example, with the ground-truth dataset, we can quantify the amount of complementary and consensus information among the multi-omics data. One potential direction is to collect data for the known cross-modality pathways (e.g., DNA methylation and gene ex-

pression pathways), which can be used to validate the consensus principle. On the other hand, the multi-omics data simulation, which naturally comes with ground truth, can serve as an alternative for validation. Although some multi-omics data simulation works have been recently developed [180, 181], they are not specifically designed to validate the interactions across modalities with 1) consensus information (e.g., co-regulation pathways), 2) complementary information (e.g., modality-specific pathways/biomarkers), and 3) endpoint irrelevant information. Thus, one promising future step is to simulate multi-omics data to validate the integration principles and corresponding methods.

One essential limitation of the current study is the relatively small sample size of the TCGA-BRCA dataset. As a purely data-driven approach, deep learning performance is significantly influenced by the amount of training data. However, we only have around 1,000 samples from the TCGA-BRCA dataset. One future direction is to test and improve our model with a larger breast cancer survival dataset or combine multi-source breast cancer survival datasets. On the other hand, besides using the TCGA database for survival analysis, we can also test the proposed integration methods' generalizability by applying them to other multi-omics datasets with various endpoints in follow-up studies.

There are multiple directions to improve the current multi-omics data integration model. One primary limitation of the current framework is the feature selection step. We apply two simple feature selection/dimension reduction methods with a focus on multi-omics data integration: unsupervised feature selection by variance ranking and unsupervised dimension reduction by PCA. One straightforward extension is to utilize more sophisticated feature selection methods, such as knowledge-guided feature selection. However, a more generalizable future direction is integrating the feature selection or dimension reduction step with our multi-modality network. Our current results have demonstrated that the feature-selection or dimension-reduction steps will impact multi-modality integration performance. To utilize the data-driven philosophy of deep learning, we can integrate the feature selection step with the feature representation step to improve model performance.

Secondly, a more sophisticated model to combine consensus learning and complementary learning may further improve multi-omics integration. The interactions among -omics modalities can be complicated so that these modalities are supposed to contain both complementary and consensus information. Thus, we propose to capture both complementary and consensus information by extending the current ConcatAE framework. Instead of using one encoder for each modality, we can use two encoders or an encoder with branches to represent both the modality-unique hidden feature and the modality-consensus feature. The modality-unique hidden features can be learned by maximizing the divergence among modalities, while the modality-consensus hidden features can be learned by minimizing the divergence among modalities. Instead of cross-modality reconstruction in CrossAE, the consensus constraints and the complementary constraints in the proposed method are both realized by divergence optimization. Hopefully, the divergence-based consensus learning can achieve better performance on higher dimension input data, which can overcome the challenge CrossAE faced, as shown in Fig. 3.12B.

Besides improving the integration framework, another future direction is to improve the survival model. This study has implemented a simple deep learning-based survival network using the negative partial log-likelihood loss. However, we did not apply any extra regularization to the survival model. One future work is to improve the survival network with regularization, such as L_1 loss on the network weights. We believe a more robust survival network will further improve the multi-omics integrated survival network.

3.2.5 Conclusions

This study has investigated two multi-modal data integration strategies: ConcatAE to integrate the complementary information among modalities and the CrossAE to integrate the consensus information among modalities. We first tested the proposed models on the simulated MNIST data. We validated the effectiveness of ConcatAE in integrating complementary information and CrossAE in integrating consensus information among multi-modality

data.

We then apply the proposed models to the multi-omics breast cancer survival data obtained from the TCGA-BRCA dataset. For the single-omics model, the miRNA expression is the most predictive for breast cancer survival analysis (0.616 ± 0.057), followed by DNA methylation and gene expression. CNV data is the least predictive for breast cancer overall survival analysis. For the multi-omics model, the ConcatAE model integrating DNA methylation and miRNA expression PCA features achieves the best performance with a C-index of 0.641 ± 0.031. The CrossAE model integrating DNA methylation and miRNA expression PCA features achieves a C-index of 0.63 ± 0.081, which also outperforms either DNA methylation or miRNA expression alone. We conclude that the DNA methylation data and miRNA expression data contain both complementary and consensus information. We can achieve improved survival analysis performance by utilizing complementary and consensus information in the integration model. As a future direction, we believe a sophisticated learning framework to integrate both consensus and complementary information in multi-omics data can improve survival analysis performance, which is essential for personalized diagnosis and treatment for breast cancer patients.

3.3 Multi-Omics Integration with Divergence-based Consensus Learning

3.3.1 Background

High-throughput multi-omics data have enabled personalized diagnosis and treatments for genetics related diseases such as cancer. However, it is extremely challenging for physicians to make sense of the molecular biomarkers directly from the -omics data tsunami. For computational scientists to integrally analyze multi-omics data, there are also a few issues to be addressed, including "the curse of dimensionality" [182] and the multi-modal analysis.

In multi-omics data integration, we aim to improve the diagnosis (e.g., cancer staging) and prognosis prediction (e.g., survival analysis) by combing the information embedded

in each modality. We categorize the multi-omics integration methods based on whether a specific class of computational models is used. Model-agnostic approaches integrate features from each modality directly, either by concatenating them after feature selection [94, 79] or integrating them after decision made of each modality [83]. In model-based approaches, models such as kernel machines, graphical models, and neural networks [80] are needed to encode the assumptions of data. Then based on the designed models, features are extracted for downstream tasks. Because multi-omics variations jointly impact disease development and patient prognosis, the good data integration methods can improve the model performance by capturing the interactions among multi-omics (e.g., gene expression pathways) compared to single-omics methods. However, many multi-omics integration studies in the literature can hardly exploit the full interactions among modalities and thus result in sub-optimal performance.

Following the previous section's investigations, because it is difficult to explore all multi-omics interactions explicitly, we investigate how to learn the interactions implicitly by deep neural networks with divergence-based regularization for classification and survival modeling tasks [183]. The novel contributions of this study include three parts:

1. We develop an effective end-to-end solution for integrating different modalities based on deep neural networks. We develop a novel regularization technique to model the observation that two modalities share consensus information about the same patient. This method can be easily extended to multiple modalities.

2. We present evidence and suggest when and how data integration should be done to integrate different modalities of data.

3. We show the effectiveness of the new divergence-based consensus regularization by extensive experiments on cancer types classification and cancer survival prediction, and by visualization of the modality-invariant features after consensus learning.

In this study, we propose to integrate multi-omics data with consensus learning (Figure 3.13). To implicitly model the interactions among modalities, we focus on maximizing

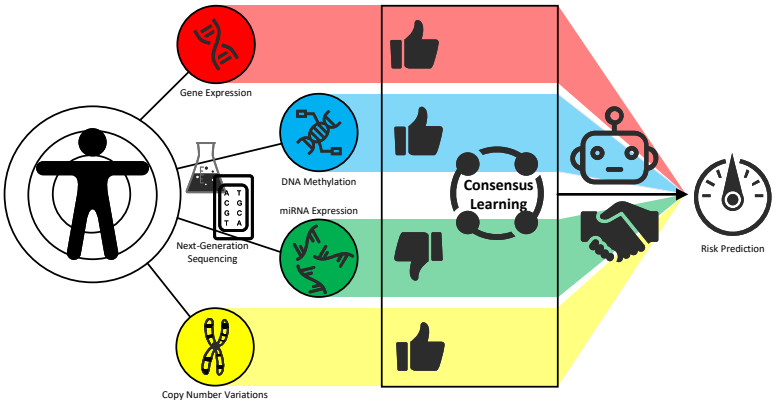


Figure 3.13: Integration of multi-omics data (e.g., gene expression, DNA methylation, miRNA expression, and copy number variations (CNVs)) with consensus learning for improved prediction performance.

the agreement between modalities. By learning modality-invariant representations with divergence-based consensus regularization, we integrate the multi-omics data in a common hidden space and improve the prediction of overall survival for breast cancer and ovarian cancer patients.

3.3.2 Materials and Methods

Datasets

For cancer types classification, we collect four TCGA cancer datasets from UCSC Xena [184] including lung adenocarcinoma (LUAD), kidney renal clear cell carcinoma (KIRC), lung squamous cell carcinoma (LUSC), and pancreatic adenocarcinoma (PAAD) (Table 3.5). Each cancer dataset consists of four -omics data with the same numbers of features, including gene expression (GeneExp), DNA methylation (DnaMeth), miRNA expression (miRNA), and copy number variations (CNVs). The GeneExp and miRNA data downloaded from UCSC Xena has already been $\log 2$ transformed. We apply min-max transformation to all four -omics data and normalize features to the range $(0, 1)$. We only keep the samples with all four -omics data for the cancer types classification. We choose two out of

Table 3.5: Overview of the TCGA samples for cancer types classification

Cancer Types	GeneExp	DneMeth	miRNA	CNVs	# of Samples with All Modalities
LUAD	585	503	564	531	454
KIRC	607	483	592	536	319
LUSC	550	412	523	503	364
PAAD	182	195	183	185	177
# of features	60,483	485,577	1,881	19,729	

Table 3.6: Overview of the TCGA samples for overall survival analysis

Cancer Types	GeneExp	DneMeth	miRNA	CNVs	# of Samples with All Modalities	# of Event	Max Survival Time	Min Survival Time
BRCA	1,222	1,234	1,207	1,106	1,061	145	8605	0
OV	379	613	499	620	362	221	5481	8

the four modalities as the input for every classification experiment configuration, and we exhaust all six combinations for our experiments.

We collect two cancer datasets directly from the TCGA data portal for overall survival analysis, including breast cancer (BRCA) and ovarian cancer (OV). Each cancer dataset also consists of the GeneExp, DnaMeth, miRNA, and CNVs data along with overall survival information (Table 3.6). The number of features is 60,483 for GeneExp, 25,978 for DnaMeth, 1,881 for miRNA, and 19,729 for CNVs in overall survival analysis. Note that we use different DnaMeth experiment data as that in the cancer types classification. We first apply $\log 2$ transformation to GeneExp and miRNA data and then perform min-max normalization to scale the features to the range (0, 1) for all -omics features. Similarly, we only keep samples with all four -omics data for overall survival analysis.

Feature Selection and Dimension Reduction A typical -omics data set usually contains only a few hundred samples but with millions of features. Thus, -omics data usually suffer from the "curse of dimensionality." Researchers typically apply various feature selection or dimension reduction techniques to the raw- omics features to mitigate the challenge. For example, Huang et al. apply co-expression analysis to reduce the number of gene expression and miRNA expression features [82]. EL-Manzalawy et al. utilize min-redundancy and max-relevance for feature selection for multi-omics data [79]. This study focuses on

integration methods and applies two simple feature selection or dimension reduction techniques as we have applied in the previous section. The first method we choose is the dimension reduction technique principal component analysis (PCA). We apply PCA to each -omics data and obtain the first n principal components(PCs) as the dimension-reduced features for integration. We will call these features PCA features hereafter. To determine the optimal numbers of PCs to keep, we apply a grid search for each task based on the sample sizes, as the number of PCs cannot exceed the training sample sizes. We apply the grid search for PCA features for cancer types classification with $n = 50, 100, 150, 200$. We do not include the number of PCs beyond 200 as the performance improvement is marginal for cancer types classification. For breast cancer overall survival prediction, we apply the grid search for PCA features with n ranges from 50 to 600 with a step size of 50. For ovarian cancer overall survival prediction, we apply the grid search for PCA features with $n = 50, 100, 150, 200$, as the sample size of ovarian cancer is smaller than breast cancer.

The second method we choose is the unsupervised univariate feature selection by variance. For each -omics data, we choose the top 1000 features with the largest variances. We will call these features high variance features hereafter. We do not apply gird search for high variance features to determine the optimal number of features as we did for PCA features for simplicity. The baseline models and proposed integration models are applied to both PCA features and high variance features.

Train Test split For both cancer types classification and overall survival analysis experiments, we perform a stratified four-fold cross-validation with 60% training, 15% validation, and 25% testing data. The cancer types classification experiments are stratified with cancer types as the label, and the overall survival analysis experiments are stratified with survival events.

Consensus Feature Representation Learning for Multi-Modality Data Integration

In this study, we propose to integrate multi-omics data with consensus constraints, aiming to generate modality-invariant representations among various -omics data. We compare the proposed network architecture's performance with the baselines, including the single-modality network and concatenation-based integration network with or without the autoencoding process. The network architectures compared in this study are visualized in Fig. 3.14. In the single-omics network (Fig. 3.14A), we use an encoder for feature representation and a decoder for reconstruction. The represented hidden features are fed into a task-specific network for classification or survival analysis.

In the concatenation-based multi-omics integration network (Fig. 3.14B), we have an encoder-decoder structure for each -omics modality, the represented hidden features are concatenated and then fed into the task-specific network (Algorithm 1). We will call this network AutoencoderConcat. We also test the performance using the concatenation-based network without the decoder part [82], which is called SimpleConcat.

Inspired by the integrative analysis of -omics data by dimension reduction to correlated structure across modalities [185], we propose to integrate the multi-omics data with consensus constraints (i.e., domain-invariant representations) as shown in Fig. 3.14C. With the encoder-decoder framework for each -omics modality, we impose a divergence-based constraint on the hidden features learned from each modality (Algorithm 2). In this study, we tested the consensus learning with the Cosine similarity [81] and Euclidean distance, respectively. For consensus learning with Cosine similarity, we maximize the Cosine similarity between hidden features learned from each -omics modality. We will refer to this framework as ConsensusCosine. We minimize the Euclidean distance between hidden features learned from each -omics modality for consensus learning with Euclidean distance. We will refer to this framework as ConsensusEuclidean. The learned hidden features from concatenation-based integration and consensus-based integration are visualized with the t-SNE plot for comparison.

For single-modality networks, concatenation-based integration networks (SimpleConcat and AutoencoderConcat), and consensus-based integration networks (ConsensusCosine and ConsensusEuclidean), we utilize the same network structures for the shared components. For simplicity, we have used three-layer fully connected networks with various numbers of neurons for encoder, decoder, and task-specific networks, respectively.

Algorithm 1: Training Concatenation-Based Classification Models

Require: Multi-modal Dataset $D = \{(X_i = (M_{i,1}, M_{i,2}), y_i)\}$
Require: $Euc(a, b)$ as the Euclidean distance between vectors a, b
Initialize modality specific encoders as Enc_1, Enc_2
Initialize modality specific encoders as Dec_1, Dec_2
Initialize classification module as CLF
while *Not Converged* **do**
  //Encoder-Decoder Training
  Sample a batch of sample from D
  $loss \leftarrow 0.$
  for *each sample* i *in the batch* **do**
    Compute reconstructed input as $\hat{M}_{i,j} = Dec_j(Enc_j(M_{i,j})$, for $j = 1, 2$
    Compute reconstruction loss $recon\_loss_i$ as $Euc(\hat{M}_{i,1}, M_{i,2}) + Euc(\hat{M}_{i,2}, M_{i,2})$
    $loss \leftarrow loss + recon\_loss_i$
  Back propagate $loss$ and update model parameters for encoders and classification module
  //Encoder-Classifier Training
  Sample a batch of sample from D
  $loss \leftarrow 0.$
  for *each sample* i *in the batch* **do**
    Compute modality specific representations as $h_{i,j} = Enc_1(M_{i,j})$, for $j = 1, 2$
    Build joint representation as $h_i = Concat(h_{i,1}, h_{i,2})$
    Compute classifiction loss clf_loss_i using y_i and prediction $\hat{y} = CLF(h_i)$
    $loss \leftarrow loss + clf\_loss_i + reg_i$
  Back propagate $loss$ and update model parameters for encoders and classification module
return *Modality specific encoders and decoders* $Enc_1, Enc_2, Dec_1, Dec_2$*, and classification module as* CLF

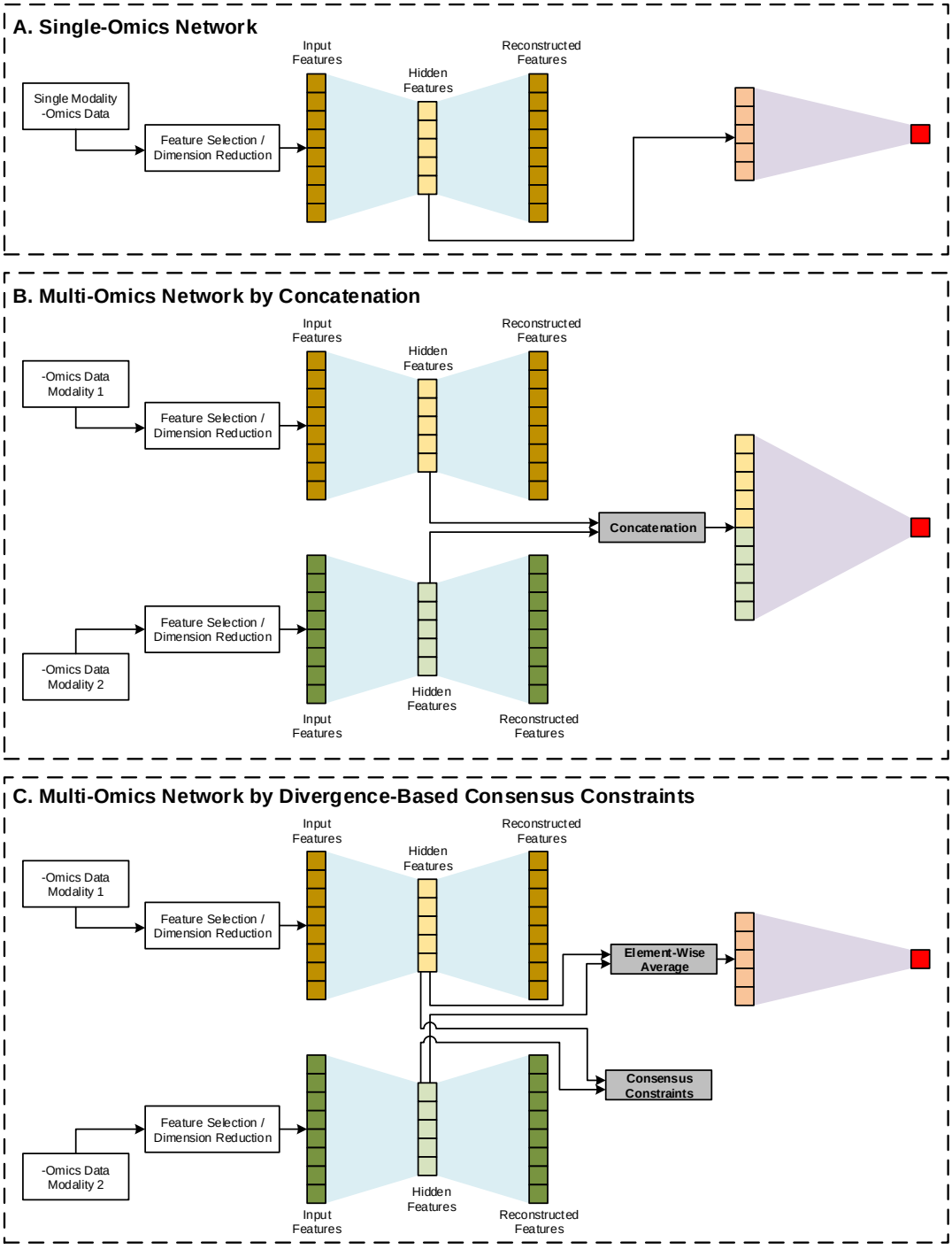


Figure 3.14: The network architectures for single-omics and multi-omics integration. A. Single-omics netowrk. B. Multi-omics network by concatenation. C. Multi-omics network by divergence-based consensus constraints

Algorithm 2: Training Consensus-Based Classification Models

Require: Multi-modal Dataset $D = \{(X_i = (M_{i,1}, M_{i,2}), y_i)\}$
Require: $Euc(a, b)$ as the Euclidean distance between vectors a, b
Initialize modality specific encoders as Enc_1, Enc_2
Initialize modality specific encoders as Dec_1, Dec_2
Initialize classification module as CLF
while *Not Converged* **do**
  //Encoder-Decoder Training
  Sample a batch of sample from D
  $loss \leftarrow 0.$
  for *each sample* i *in the batch* **do**
    Compute reconstructed input as $\hat{M}_{i,j} = Dec_j(Enc_j(M_{i,j})$, for $j = 1, 2$
    Compute reconstruction loss $recon\_loss_i$ as $Euc(\hat{M}_{i,1}, M_{i,2}) + Euc(\hat{M}_{i,2}, M_{i,2})$
    $loss \leftarrow loss + recon\_loss_i$
  Back propagate $loss$ and update model parameters for encoders and classification module
  //Encoder-Classifier Training
  Sample a batch of sample from D
  $loss \leftarrow 0.$
  for *each sample* i *in the batch* **do**
    Compute modality specific representations as $h_{i,j} = Enc_1(M_{i,j})$, for $j = 1, 2$
    Build joint representation as $h_i = \frac{1}{2}(h_{i,1} + h_{i,2})$
    Compute classifiction loss clf_loss_i using y_i and prediction $\hat{y} = CLF(h_i)$
    Compute consensus regularization, for example, the Euclidean-based regularization as $reg_i = Euc(h_{i,1}, h_{i,2})$
    $loss \leftarrow loss + clf\_loss_i + reg_i$
  Back propagate $loss$ and update model parameters for encoders and classification module
return *Modality specific encoders and decoders* $Enc_1, Enc_2, Dec_1, Dec_2$*, and classification module as* CLF

Endpoint 1: Cancer Types Classification

The first endpoint for our proposed multi-omics integration network is multi-class classification. For this endpoint, we use a fully-connected network for classification and trained with cross-entropy loss. The classification performance is evaluated by accuracy, precision, recall, and area under the curve (AUC) for binary classification. For multi-class classification, we use accuracy, weighted precision, and weighted recall as the evaluation metrics. These metrics are in the range of $[0, 1]$, and the higher, the better.

Endpoint 2: Survival Risk Analysis

Survival analysis aims to predict the expected duration of time until one or more events happen by modeling the time to event data. The proportional hazards model assumes the covariates are multiplicatively related to the hazard [173]. Assuming the proportional hazards assumption holds, the Cox proportional hazards model can estimate the effect parameters without considering the hazard function [169]. Thus, the Cox proportional hazards model is semi-parametric. Let $X_i = X_{i1}, ..., X_{ip}$ be covariates for subject i. The hazard function for the Cox proportional hazards model has the form:

$$\lambda(t|X_i) = \lambda_0(t) \exp(\beta_1 X_{i1} + ... + \beta_p X_{ip}) = \lambda_0(t) \exp(X_i \cdot \beta) \tag{3.7}$$

This expression gives hazards function at time t for subject i with covariate vector X_i. The likelihood of the event to be observed occurring for subject i at time Y_i can be written as:

$$L_i(\beta) = \frac{\lambda(Y_i|X_i)}{\sum_{j:Y_j \geq Y_i} \lambda(Y_i|X_j)} = \frac{\lambda_0(Y_i)\theta_i}{\sum_{j:Y_j \geq Y_i} \lambda_0(Y_i)\theta_j} = \frac{\theta_i}{\sum_{j:Y_j \geq Y_i} \theta_j} \tag{3.8}$$

where $\theta_j = \exp(X_j \cdot \beta)$ and the summation is over the set of subjects j where the event has not occurred before time Y_i (including subject i itself). $L_i(\beta)$ is called a partial likelihood as the effect of the joint probability can be estimated without modeling the change

of the hazard over time. Obviously $0 < L_i(\beta) < 1$. Treating the subjects as if they were statistically independent of each other, we can obtain the joint probability of all realized events:

$$L(\beta) = \prod_{i:C_i=1} L_i(\beta) \tag{3.9}$$

where the occurrence of the event is indicated by $C_i = 1$. The corresponding log partial likelihood is

$$l(\beta) = \sum_{i:C_i=1} (X_i \cdot \beta - \log \sum_{j:Y_j \geq Y_i} \theta_j) \tag{3.10}$$

With deep learning development, the Cox proportional hazards model has been extended with deep neural networks. Cox-Time [174] and Deep Surv [175] similarly replacing the linear relationship $\exp(X_i \cdot \beta)$ with non-linear transformation as $\exp(f_\phi(X_i) \cdot \beta)$, where f_ϕ is a neural network parametrized by ϕ, for example, a fully connected neural network. In addition, the authors also proposed L_1 and L_2 regularization terms, respectively, on the parameter ϕ to reduce over-fitting of the models. This study uses a fully connected neural network for the Cox proportional hazards model with the log partial likelihood as the loss function.

To evaluate the risk scores produced by survival models, researchers have developed various metrics to measure the concordance between the predicted risk scores and the actual survival time. Following the previous studies in deep-learning-based survival analysis [82], we evaluate the overall survival analysis performance with the concordance index (C-index) [178]. C-index evaluates how well the survival risk we computed aligns with the actual survival time, i.e., for two individuals $(X_1, T_1), (X_2, T_2)$, C-index $= \Pr(\lambda(X_2) > \lambda(X_1) | T_2 > T_1)$. We also evaluate the survival analysis with the Kaplan-Meier curve and the log-rank test.

Implementation and Experiments

The train-test split for cross-validation, the classification metrics, the t-SNE visualization are implemented with [179]. The Kaplan-Meier plot and log-rank test are performed with lifelines [186]. The neural networks are designed and implemented with PyTorch 1.1.0. For cancer types classification, we use a batch size of 8, Adam optimizer with a learning rate of 0.001, and training epochs of 200. The test performance is based on the models with the best validation performances. For survival analysis, we use a batch size of 128, Adam optimizer with a learning rate of 0.001, and training epochs of 200. As we observed an over-fitting for models with the best validation performance, the test performance is based on the models until the end of training (200 epochs). More details of the model implementation and training details can be found at Github repo (`https://github.com/tongli1210/TCGA_Omics`).

3.3.3 Results

Cancer Types Classification

We perform cancer types classification using four TCGA datasets (LUAD, KIRC, LUSC, and PAAD) with six binary classifications and one four-class classification. The binary classification is evaluated with accuracy, precision, recall, and AUC. The four-class classification is evaluated with accuracy, precision, and recall. For each cancer-types classification task, we perform PCA feature reduction with a grid search of the number of PCs from 50 to 200 with a step size of 50. The percentage of variances explained by the PCs are compared for each classification task and train-test split. Based on the results, we observe that the first 50 PCs can explain about 70%, 90%, 80%, 70% of variances for GeneExp, DnaMeth, miRNA, and CNV, respectively.

We perform cancer types classification using two of the proposed frameworks AutoencoderConcat and ConsensusEuclidean. The classification results using various numbers of

PCA features and high variance features are compared. Furthermore, we visualize the accuracy of integration-based cancer types classification for some selected experiments using radar plots in Fig. 3.15. For single-omics binary classification, GeneExp, DnaMeth, and miRNA achieve similar performance and significantly outperform CNV. For multi-omics binary classification, the accuracy for both integration frameworks AutoencoderConcat or ConsensusEuclidean is very high except for LUAD vs. LUSC. Since LUAD and LUSC are both lung cancers, we suspect the separation of LUAD vs. LUSC is harder than the other binary classification tasks, which also explains the performance drop for four-class classification that includes both LUAD and LUSC. Comparing the two multi-omics integration approaches, we find that the AutoencoderConcat consistently outperforms the ConsensusEuclidean integration when using PCA features. In contrast, the AutoencoderConcat integration consistently outperforms the ConsensusEuclidean integration when using high variance features (Fig. 3.15).

Cancer Overall Survival Analysis

Evaluation of Overall Survival Analysis using C-index Like the cancer types classification task, we perform PCA feature reduction with a grid search of the number of PCs for breast cancer and ovarian cancer overall survival prediction. The percentages of variances explained by the PCs are compared for various modalities and train-test splits for breast cancer. Based on the results, we observe that the first 50 PCs can explain about 72%, 92%, 82%, 64% of variances for GeneExp, DnaMeth, miRNA, and CNV, respectively. The percentages of variances explained by the PCs are compared for various modalities and train-test splits for ovarian cancer. Based on the results, we observe that the first 50 PCs can explain about 72%, 91%, 84%, 71% of variances for GeneExp, DnaMeth, miRNA, and CNV, respectively.

We evaluate the overall survival analysis performance with PCA features and high variance features using C-index for both single-omics and multi-omics approaches. For breast

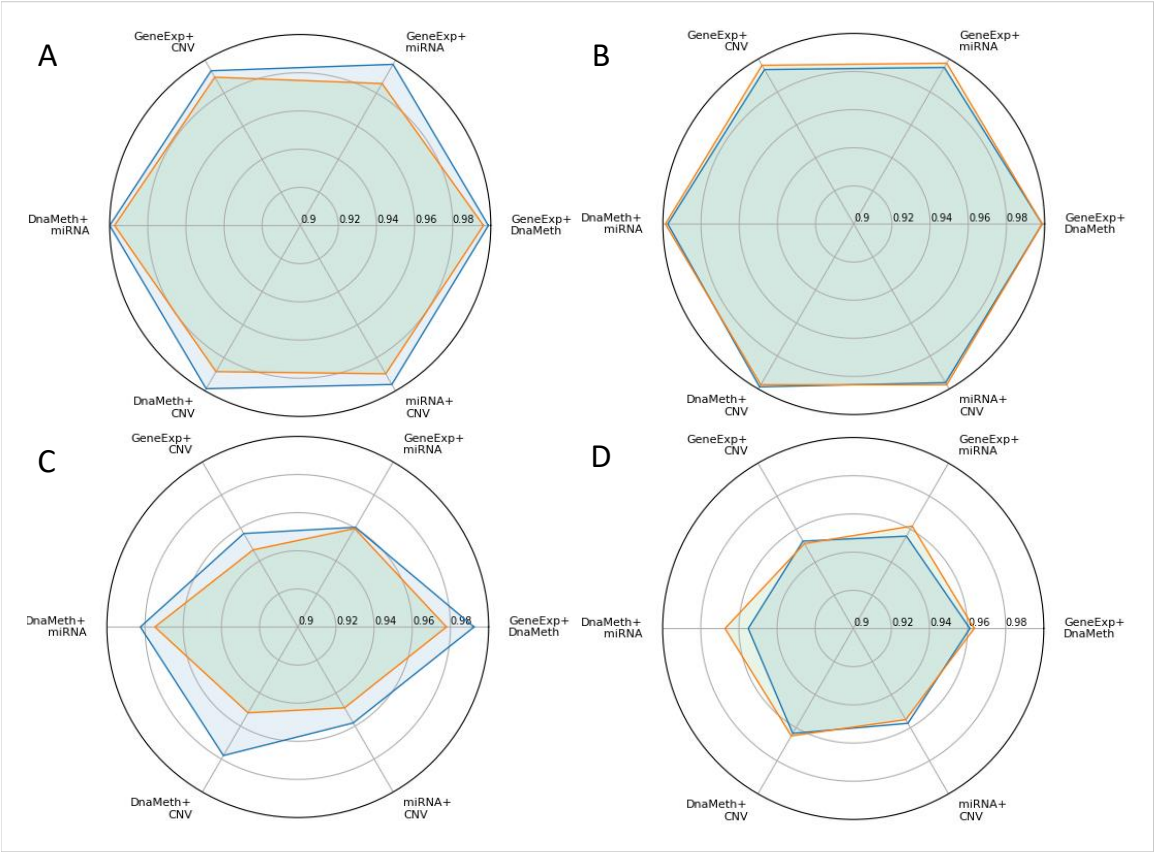


Figure 3.15: Radar plots of the accuracy for cancer types classification using AutoencoderConcat framework and ConsensusEuclidean framework for integration. A. LUAD vs. KIRC binary classification using the top 100 PCA features. B. LUAD vs. KIRC binary classification using the high variance features. C. LUAD vs. KIRC vs. LUSC vs. PAAD four-class classification using the top PCA features. D. LUAD vs. KIRC vs. LUSC vs. PAAD four-class classification using the high variance features. Blue lines: AutoencoderConcat. Yellow lines: ConsensusEuclidean.

Table 3.7: BRCA Overall Survival (OS) Analysis C-Index with Single-Omics

Features	GE	DM	mR	CNV
PCA 300	0.585±0.065	0.591±0.064	0.629±0.089	0.496±0.051
High Variance	0.529±0.033	0.581±0.066	0.614±0.041	0.503±0.071

cancer (BRCA) survival analysis, the best performance for PCA features is achieved with 300 PCs. Thus, we select the results using 300 PCs in Table 3.7 for single-omics experiments and Table 3.8 for multi-omics experiments. For ovarian cancer (OV) survival analysis, the best performance for PCA features is achieved with 50 PCs. Thus, we select the results using 50 PCs in Table 3.9 for single-omics experiments and Table 3.10 for multi-omics experiments.

From the tables, we can observe that overall survival analysis's performance improves after integrating multi-omics data with either PCA features or high variance features. For BRCA survival analysis, the concatenation-based integration outperforms consensus-based integration in some -omics combinations while consensus-based integration outperforms concatenation-based integration in the other -omics combinations. The consensus-based integration ConsensusCosine using PCA features of DnaMeth and miRNA achieves the best performance (0.671 ± 0.046). The consensus-based integration ConsensusEuclidean using PCA features of miRNA and CNV achieves a similar performance (0.667 ± 0.073).

For OV survival analysis, we also observed that the concatenation-based integration outperforms consensus-based integration in some -omics combinations while consensus-based integration outperforms concatenation-based integration in the other -omics combinations. The concatenation-based integration AutoencoderConcat using PCA features of miRNA and CNV achieves the best performance (0.571 ± 0.036). We have also compared the C-index of concatenation-based and consensus-based integration methods using radar plot (Fig. 3.16).

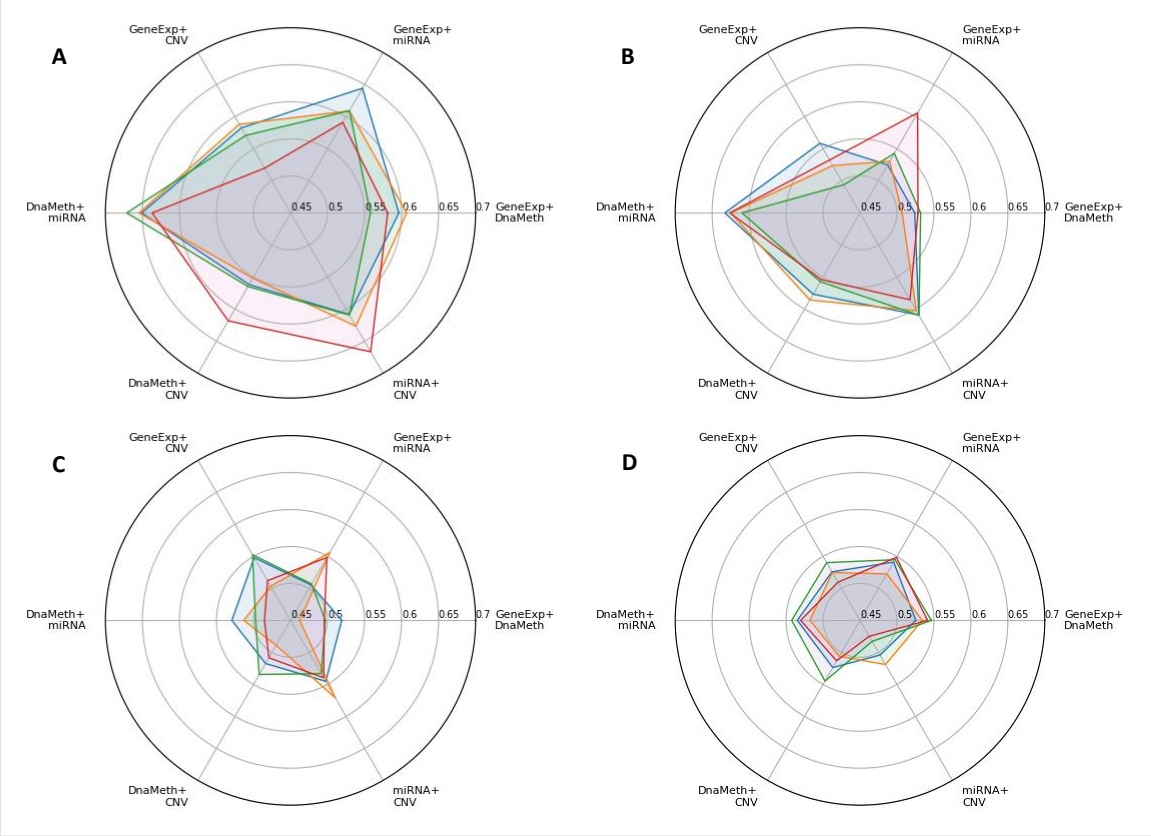


Figure 3.16: Radar plot of the C-Index for overall survival analysis using concatenation and consensus based multi-omics integration. A. Breast cancer (BRCA) overall survival (OS) C-index with Top 300 PCA features. B. Breast cancer (BRCA) overall survival (OS) C-index with high variance features. C. Ovarian cancer (OV) overall survival (OS) C-index with Top 50 PCA features. D. Ovarian cancer (OV) overall survival (OS) C-index with high variance features. Green lines: SimpleConcat. Blue lines: AutoencoderConcat. Magenta lines: ConsensusCosine. Yellow lines: ConsensusEuclidean.

Table 3.8: BRCA Overall Survival (OS) Analysis C-Index with Multi-Omics

Features	Methods	GE+DM	GE+mR	GE+CNV	DM+mR	DM+CNV	mR+CNV
PCA 300	SimpleConcat	0.596±0.054	**0.644±0.057**	0.583±0.066	0.651±0.055	0.562±0.057	0.608±0.039
	AutoencoderConcat	**0.607±0.073**	0.609±0.053	**0.588±0.105**	0.655±0.062	0.552±0.061	0.627±0.062
	ConsensusCosine	0.558±0.081	0.609±0.078	0.571±0.054	**0.671±0.046**	0.565±0.024	0.609±0.056
	ConsensusEuclidean	0.581±0.057	0.591±0.115	0.52±0.025	0.637±0.073	**0.618±0.086**	**0.667±0.073**
High Variance	SimpleConcat	0.524±0.024	0.525±0.04	**0.558±0.019**	**0.633±0.042**	0.577±0.04	0.609±0.054
	AutoencoderConcat	0.507±0.036	0.53±0.052	0.524±0.038	0.625±0.023	**0.586±0.068**	0.603±0.04
	ConsensusCosine	**0.532±0.017**	0.543±0.011	0.494±0.052	0.61±0.068	0.557±0.034	**0.61±0.056**
	ConsensusEuclidean	0.528±0.018	**0.606±0.041**	0.532±0.02	0.626±0.056	0.554±0.024	0.585±0.049

Table 3.9: OV Overall Survival (OS) Analysis C-Index with Single-Omics

Features	GE	DM	mR	CNV
PCA 50	0.504±0.016	0.488±0.046	0.549±0.037	0.523±0.037
High Variance	0.49±0.024	0.524±0.026	0.506±0.046	0.497±0.03

Visualization of Risk Prediction with Kaplan-Meier Plot

Besides the C-index for evaluating survival analysis, we also perform the Kaplan-Meier plot with the log-rank test to visualize the hazard prediction. We group the testing samples into the low-risk and high-risk groups by the median of all predicted hazards. If a testing sample's predicted hazard is lower than the median hazard of all testing samples, it will be assigned to the low-risk group; otherwise, it will be assigned to the high-risk group. We visualize the two groups with Kaplan-Meier curves and test the separation of these two groups with the log-rank test. Fig. 3.17 presents the Kaplan-Meier plot of breast cancer(BRCA) overall survival (OS) analysis using the integration of the top 300 PCA features of DnaMeth and miRNA. From this Kaplan-Meier plot, we observe that all four integration methods are able to achieve a good separation of the low-risk and high-risk groups. Fig. 3.18 presents the Kaplan-Meier plot of breast cancer (BRCA) overall survival (OS)

Table 3.10: OV Overall Survival (OS) Analysis C-Index with Multi-Omics

Features	Methods	GE+DM	GE+mR	GE+CNV	DM+mR	DM+CNV	mR+CNV
PCA 50	SimpleConcat	**0.519±0.034**	0.505±0.033	0.548±0.015	**0.529±0.075**	0.517±0.044	0.545±0.057
	AutoencoderConcat	0.462±0.027	**0.556±0.059**	0.503±0.027	0.513±0.014	0.489±0.074	**0.571±0.036**
	ConsensusCosine	0.497±0.011	0.507±0.043	**0.553±0.053**	0.497±0.053	**0.535±0.037**	0.533±0.036
	ConsensusEuclidean	0.496±0.039	0.548±0.025	0.512±0.026	0.485±0.047	0.508±0.026	0.54±0.067
High Variance	SimpleConcat	0.525±0.034	0.541±0.053	0.526±0.036	0.535±0.017	0.524±0.021	0.504±0.051
	AutoencoderConcat	0.534±0.032	0.523±0.024	0.525±0.016	0.518±0.03	0.506±0.031	**0.519±0.058**
	ConsensusCosine	**0.547±0.045**	0.545±0.027	**0.54±0.052**	**0.542±0.015**	**0.545±0.068**	0.483±0.035
	ConsensusEuclidean	0.541±0.053	**0.549±0.053**	0.51±0.066	0.53±0.038	0.513±0.073	0.475±0.032

analysis using the integration of the top 300 PCA features of miRNA and CNV. From this Kaplan-Meier plot, we can observe the ConsensusEuclidean achieves the best low-risk and high-risk group separation, while the SimpleConcat has the worst low-risk and high-risk group separation. Additional Kaplan-Meier plots can be found in [183].

Visualization of the Hidden Features with t-SNE and Scatter Plot

We use t-SNE to visualize the hidden features learned by four multi-omics integration methods (concatenation-based: SimpleConcat and AutoencoderConcat; consensus-based: ConsensusCosine and ConsensusEuclidean). After reducing the dimensions of the hidden features to two, we visualize the decomposed hidden features with scatter plots (Fig. 3.19. We illustrate the integration of DnaMeth and miRNA in Fig. 3.19A and the integration of miRNA and CNV in Fig. 3.19B. From the scatter plots, we observe that the multi-omics hidden features learned by consensus-based integration align better compared to that of concatenation-based integration. The results demonstrate that the Cosine-similarity-based consensus learning (ConsensusCosine) and the Euclidean-based consensus learning (ConsensusEuclidean) improve the cross-modality alignment for multi-omics data, which contributes the performance improvements on survival analysis.

3.3.4 Discussions

If using 1,000 high variance features in the cancer types classification, consensus-based integration consistently outperforms concatenation-based integration. While if using top 100 PCA features, concatenation-based integration consistently outperforms the consensus-based integration. Because the consensus-based integration is based on the mutual information co-existing in multiple -omics modalities, when features are more compressed as in PCA features, the concatenation strategy works better. Thus, mutual information among modalities is essential when applying consensus-based integration. For example, if integrating two distinct modalities of a patient (e.g., gene expression vs. MRI imag-

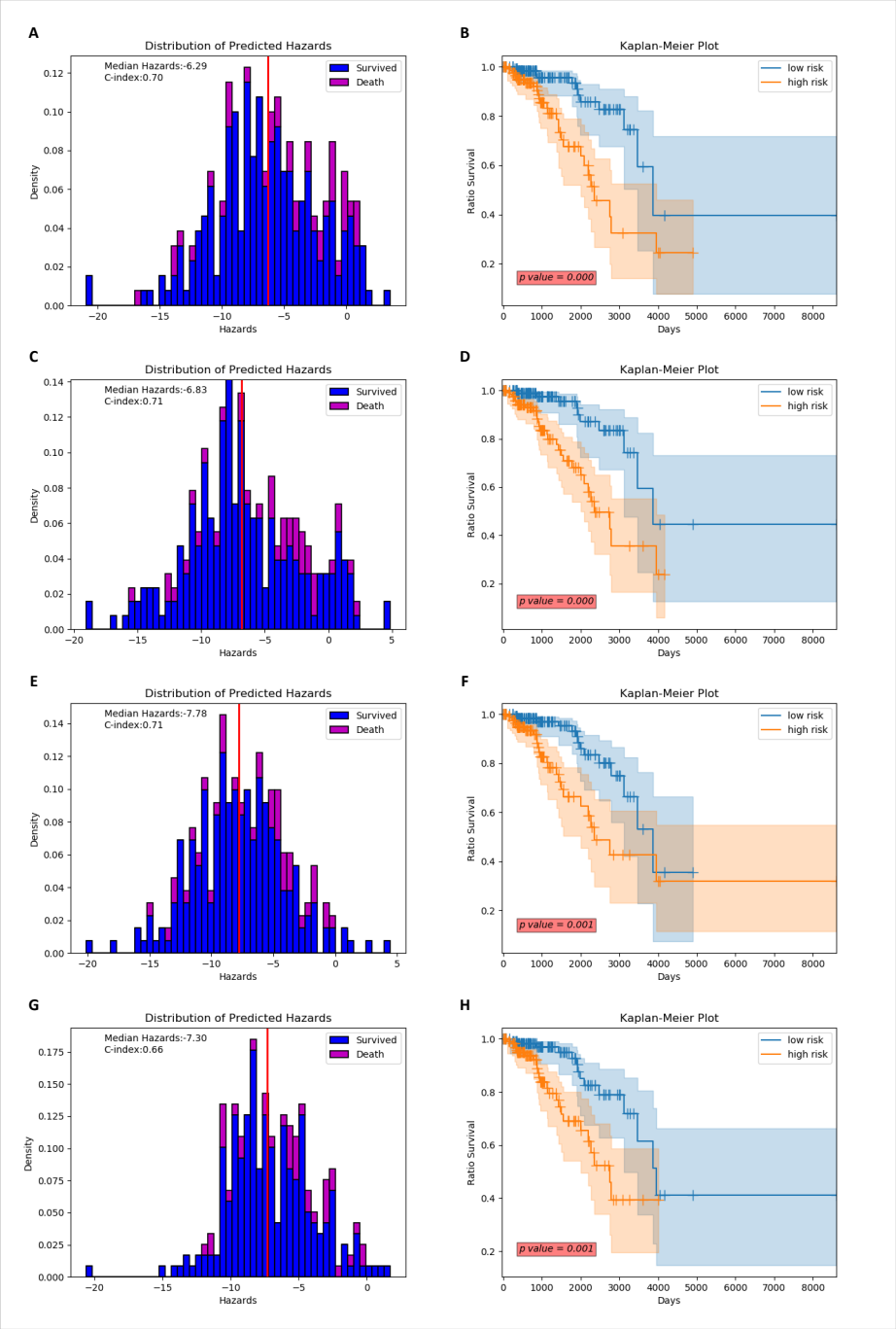


Figure 3.17: Kaplan-Meier plot of the breast cancer (BRCA) overall survival (OS) prediction by integrating the top 300 PCA features of DnaMeth and miRNA on train-test split fold 1. A. Distribution of predicted hazards with SimpleConcat. B. Kaplan-Meier plot of OS prediction with SimpleConcat. C. Distribution of predicted hazards with AutoencoderConcat. D. Kaplan-Meier plot of OS prediction with AutoencoderConcat. E. Distribution of predicted hazards with ConsensusCosine. F. Kaplan-Meier plot of OS prediction with ConsensusCosine. G. Distribution of predicted hazards with ConsensusEuclidean. H. Kaplan-Meier plot of OS prediction with ConsensusEuclidean.

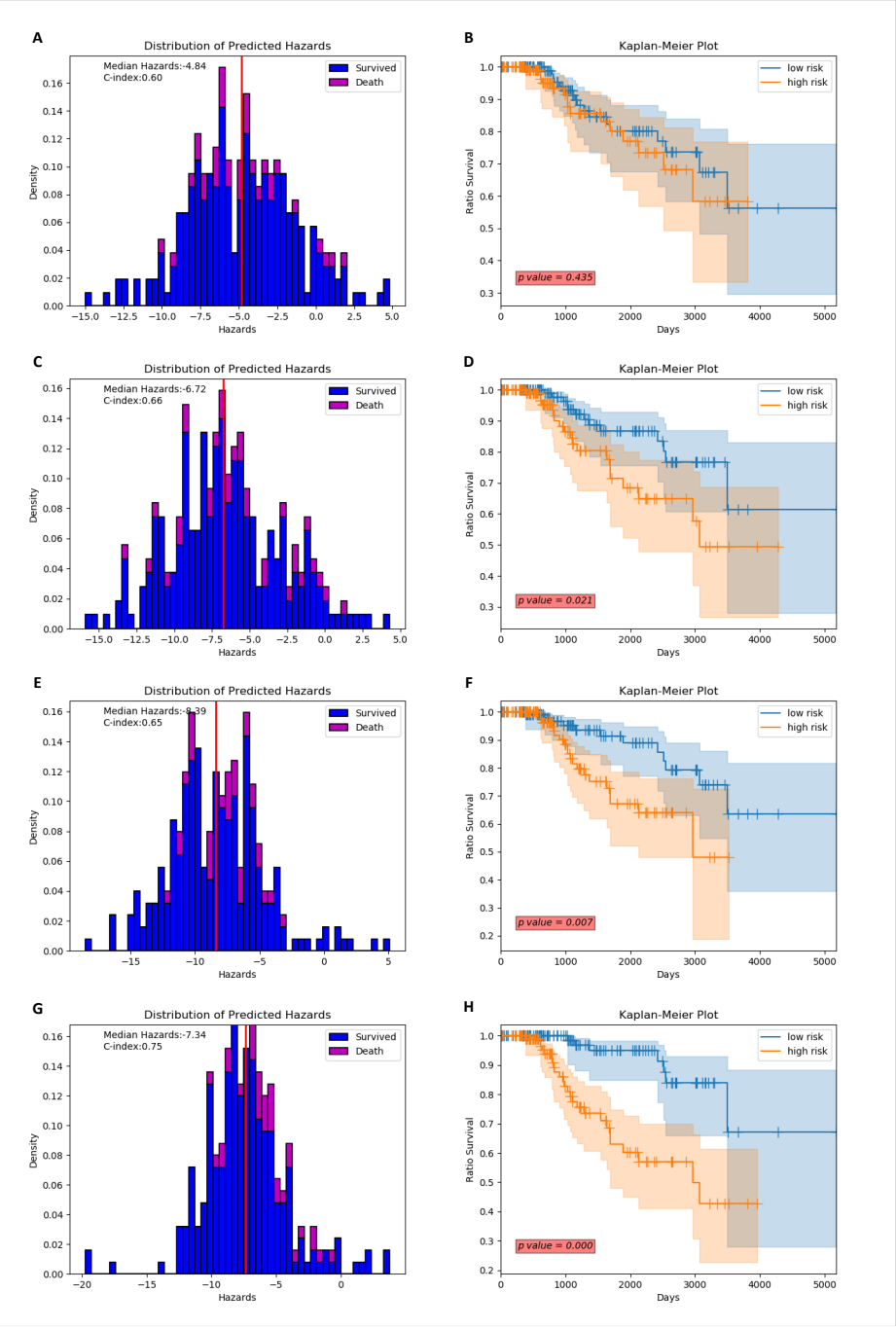


Figure 3.18: Kaplan-Meier plot of the breast cancer(BRCA) overall survival (OS) prediction by integrating the top 300 PCA features of miRNA and CNV on train-test split fold 2. A. Distribution of predicted hazards with SimpleConcat. B. Kaplan-Meier plot of OS prediction with SimpleConcat. C. Distribution of predicted hazards with AutoencoderConcat. D. Kaplan-Meier plot of OS prediction with AutoencoderConcat. E. Distribution of predicted hazards with ConsensusCosine. F. Kaplan-Meier plot of OS prediction with ConsensusCosine. G. Distribution of predicted hazards with ConsensusEuclidean. H. Kaplan-Meier plot of OS prediction with ConsensusEuclidean.

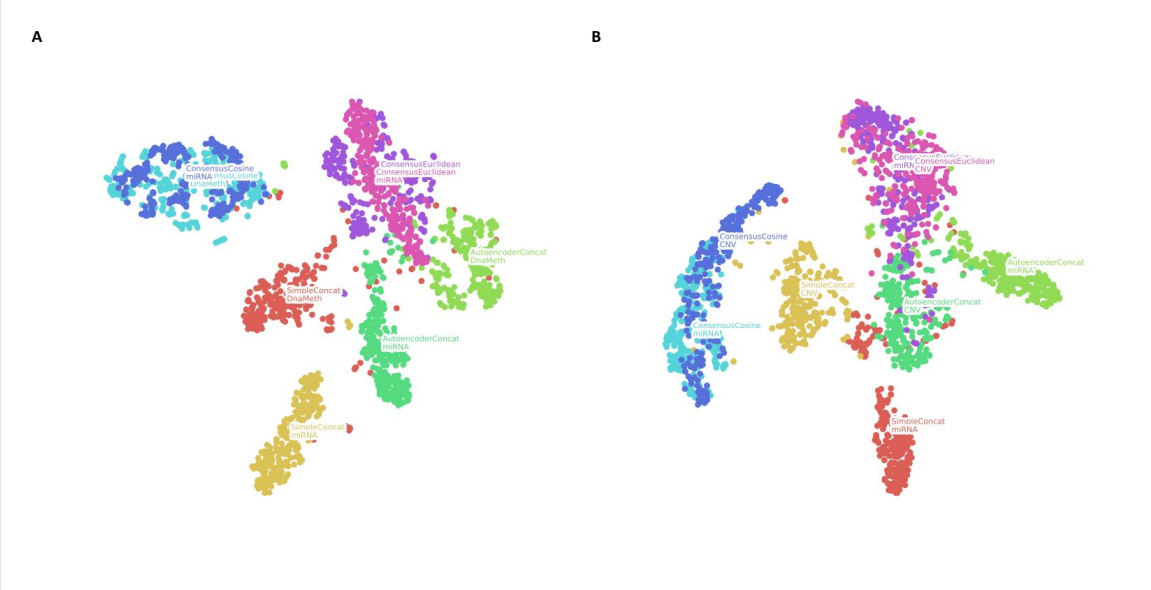


Figure 3.19: Scatter Plot of the hidden features generated for breast cancer (BRCA) overall survival (OS) analysis after t-SNE decomposition. A. t-SNE scatter plot for hidden features represented from the top 300 PCA features of DnaMeth and miRNA (cross-validation fold 1). B. t-SNE scatter plot for hidden features represented from the top 300 PCA features of miRNA and CNV (cross-validation fold 2).

ing), consensus learning may not be a good option. In overall survival prediction, both the concatenation-based and consensus-based integration models consistently outperform the single-modality models. However, the rank of their performance varies with different choices of modality combinations. A future study is needed to understand when consensus works better than concatenation in survival prediction.

In this study, we have developed a multi-omics data integration method by divergence-based modality-invariant representation. To impose consensus constraints among modalities, we maximize the Cosine similarity (ConsensusCosine) or minimize the Euclidean distances (ConsensusEucldiean) on the features represented from each -omics modality. The first future direction is to replace the Cosine similarity or Euclidean distance with other divergence criteria such as Wasserstein distance or adopt adversarial learning for modality-invariant representation instead of the predefined divergence metrics. That is to differentiate features represented from various modalities by training the discriminator and

the modality-specific encoders in an adversarial fashion.

The second future direction is to integrate the feature selection or dimension reduction step with our multi-modality network. We have currently utilized variance for feature selection and PCA for dimension reduction, respectively, which have shown influences on multi-modality integration performance. As deep learning can offer fully data-driven analysis, we can integrate the feature selection step with the modality-invariant representation step to improve the overall model performance. However, to enable such automatic feature engineering, one bottleneck could be the amount of data available for training.

Another future direction is to utilize multi-task learning with multi-view learning. In this study, we train encoders for each -omics modality (e.g., GeneExp, DnaMeth, miRNA, and CNV) independently when applying to various endpoints (e.g., cancer types classification and survival analysis). However, for encoders of a specific modality (e.g., GeneExp), they transform raw -omics features to hidden spaces with the same dimensions, despite that hidden features might be used for various cancer types (BRCA vs. OV) and prediction tasks (e.g., cancer types classification vs. survival analysis). Thus, we will apply multi-mask learning to combine the training of these encoders and obtain potential performance improvements.

3.3.5 Conclusions

This study presented an effective end-to-end solution for integrating multi-omics data based on deep neural networks. We developed a new divergence-based consensus regularization techniques to capture the consensus information among modalities to improve prediction performance. Although we only tested this new algorithm for a two-modality integration scenario with two divergence metrics (i.e., Cosine similarity and Euclidean distance), this novel method can be easily extended to multiple modalities and advanced divergence metrics. Through our analysis, we demonstrated why and how to integrate different modalities of data. Our experiment results validated the effectiveness of our new method for both can-

cer types classification and survival prediction. In the visualization of features, we observed that our approach could identify more consensus among modalities with clearer separation margins.

CHAPTER 4

SEMI-SUPERVISED LEARNING FOR MEDICAL IMAGING INFORMATICS

4.1 Introduction

The computer-aided diagnosis (CAD) system utilizes digital imaging processing and machine learning for medical imaging analytics. CAD systems aim to help doctors interpret medical images to speed up the diagnosis and reduce human biases. A typical CAD system consists of image quality control, feature extraction, predictive modeling, and model visualization, which enables automatic decision making with reliable and reproducible performance [35]. With the ability to quantitatively represent medical images, the CAD system helps detect rare events and subtle changes that may be extremely challenging for human observers. Thus, researchers have developed CAD systems for multiple imaging modalities including CT [187], MRI [188], and whole-slide images [189], which have significantly improved the diagnosis of various diseases.

Optical Endomicroscopy (OE) is a newly emerged endoscopic imaging based on confocal microscopy, spectroscopy-based imaging, or optical coherence tomography (OCT) [190]. By combining endoscopy and microscopy, OE can function as 'optical biopsies,' which enables real-time in situ biopsy instead of the conventional biopsy and histopathology. Using an OE technique, the physician can make real-time clinical decisions about the grade of dysplasia, if present, and potentially treat the patient during the same endoscopic session. Thus, this novel technique can significantly decrease the waiting time from the time of diagnosis (TOD) to time for endoscopic treatment (TET). The newly emerged OE technique provides gastrointestinal endoscopists the opportunity to evaluate the esophageal lining and mucosa in real-time through optical biopsy. Compared with the conventional biopsy through endoscope plus microscopic examination afterward, OE can improve clini-

cal care for patients with gastrointestinal conditions like Barret's esophagus (BE). BE is a disorder defined as abnormal changes from normal squamous epithelium to the columnar epithelium. The abnormal changes usually happen in the lower portion of the esophagus [191]. BE is a well-known risk factor for esophageal adenocarcinoma (EAC). Once BE is diagnosed, doctors will perform a biopsy and use the histologic severity of the targeted BE tissue to determine the cancer prevention surveillance intervals and treatment recommendations. With the limited amount of tissue collected during a biopsy and the time delay between biopsy and actual diagnosis, OE is a promising novel technique for real-time diagnosis using optical biopsy. Clinical trials have demonstrated that OE could achieve improved clinical care quality for patients with BE [192, 193, 194]. However, a large number of microscopic images are generated during an OE session. Human examination of all microscopic images in real-time through an OE session can be demanding and prone to errors. Thus, a fast and reliable CAD system to automatically process these microscopic images is essential to enable real-time diagnosis for BE patients using OE techniques.

To build a reliable CAD system, we typically need substantial labeled training data to select the optimal feature subsets and train robust classification models through supervised learning. However, accumulating a large labeled dataset for OE images can be expensive and time-consuming, and there are no public OE datasets available yet. On the other hand, it is relatively easy to collect a significant number of unlabeled images through each OE session. Thus, considering the lack of labeled OE images and the easy access to unlabeled OE images, we propose to improve the classification of OE images by utilizing the unlabeled images through semi-supervised learning. In the previous study, we have applied handcrafted feature extraction and label propagation methods for semi-supervised learning [195]. With the rapid development of deep learning and their massive success in natural image processing, we propose to improve our previous method by exploiting the convolutional autoencoders (CAE) and developing a semi-supervised deep neural network, CAESNet, for improving the multi-class classification of BE. With extensive experiments on the

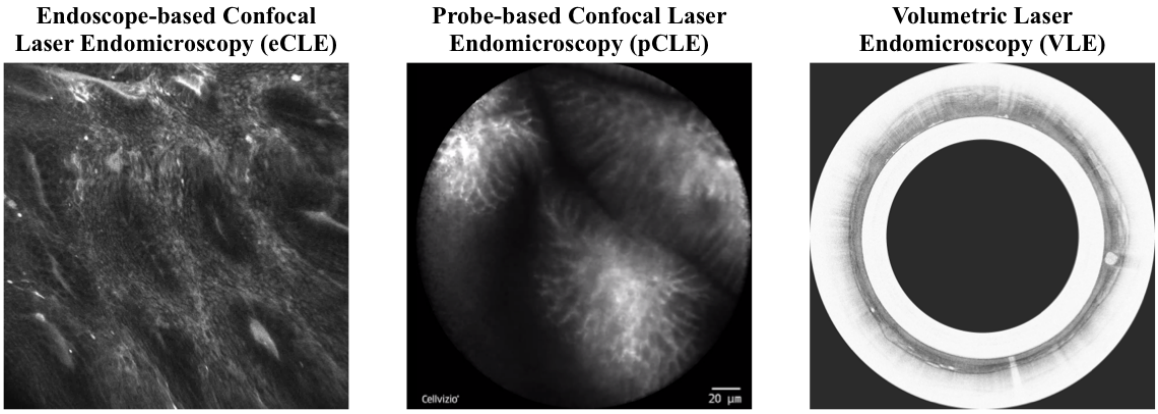


Figure 4.1: Example Images of three types of commercial OE systems. Left: Endoscope-based confocal laser endomicroscopy (eCLE). Middle: Probe-based confocal laser endomicroscopy (pCLE). Right: Volumetric laser endomicroscopy (VLE).

OE dataset collected at Emory University, we have demonstrated the superior performance of our semi-supervised CAESNet compared with all baselines [196].

4.2 Related Works

4.2.1 Optical Endomicroscopy

Optical endomicroscopy (OE) is a novel optical technology integrating endoscopy with microscopy for in situ diagnosis. It enables real-time diagnosis and treatment, in contrast to the delay in treatment due to the inherent time needed to obtain a final diagnosis by histopathology.

Currently, three types of commercial OE systems are available for clinical use: endoscope-based confocal laser endomicroscopy (eCLE), probe-based confocal laser endomicroscopy (pCLE), and volumetric laser endomicroscopy (VLE) (Figure 4.1) [190], which have been developed and approved in 2004, 2006, and 2013 respectively. Table 4.1 compares specifications of the three technologies. They are significantly different in resolution, acquisition position, acquisition speed, and microscopic imaging presentation, and we refer the readers to these papers for detailed comparison [190, 197, 198, 199]. As clinical trials suggest that eCLE has better performance than pCLE in the diagnosis of esophageal diseases [200], we

Table 4.1: Comparison of Specifications for Three OE Technologies

	eCLE	**pCLE**	**VLE**
Company	Pentax, Tokyo, Japan, and Optiscan, Victoria, Australia	Mauna Kea Technologies, Paris, France	NinePoint Medical, Cambridge, MA
Model	Pentax ISC 1000		Nvision VLE
Data Format	Surface Images	Surface image video	Helical scan video
Axial Information	Multiple z-planes	Single z-plane	Multiple z-planes
Axial Resolution	7 μm	NA	7 μm
Lateral Resolution	0.7 μm	3.5 μm	NA
Z-depth	0 – 250 μm	60 μm	0 – 3000 μm
Format Size	500×500	600 μm diameter	6 cm scan acquisition length
Image Size	1024×1024	580×576	NA
Speed	0.8-1.2 frame/sec	12 frame/sec	1200 slices / 90 sec
Intravenous Fluorescein Required	Y	Y	N

focus on developing a semi-supervised classification method for eCLE images. We believe this method can be readily generalized to other OE modalities.

4.2.2 Barret's Esophagus

Barrett esophagus (BE), a well-known risk factor for esophageal adenocarcinoma (EAC) [201], is characterized by the abnormal changes from normal squamous epithelium to the columnar epithelium at the lower portion of the esophagus. The population incidence of EAC in the United States has an estimated rise of 300% to 350% since the 1970s [202, 203], making up 60% of the new esophageal cancer diagnosis in the U.S. in 2009 [204], with only a 5-year survival rate of 15 to 20% [205]. The progression from BE to adenocarcinoma involves multiple stages: nondysplastic BE, low-grade dysplasia (LGD), high-grade dysplasia (HGD), and finally adenocarcinoma [206].

The impact of BE on the mortality from EAC is still unclear, which makes directly screening for BE controversial [207]. Thus, physicians identify patients with BE through either subjective selectively screening [208] or an upper endoscopy performed for an unrelated reason. The rapid development of OE may provide another mechanism to help detect the neoplastic changes earlier than conventional endoscopy and improve the efficiency of cancer surveillance [200]].

4.2.3 Computer-Aided Diagnosis for BE Classification

The CAD-based BE classification using OE has become an emerging field of research due to the rapidly increasing population incidence of EAC. It can realize real-time diagnosis, improve the patients' prognosis by early detection, and reduce the physicians' workload. Conventional CAD pipeline includes image quality control, feature extraction, predictive modeling, and model visualization. For example, Grisan et al. applied a support vector machine to identify gastric metaplasia vs. intestinal metaplasia using rotation invariant local binary pattern features [209]. Veronese et al. improved the approach via a two-stage classification pipeline [210], and Ghatwary et al. applied image enhancement before feature extraction for increased overall accuracy [211].

More recently, deep learning has also been applied to the classification of BE using OE images. For example, Mendel et al. have implemented deep convolutional neural networks (CNN) for binary classification of patients into EAC and non-EAC, realizing a sensitivity of 0.94 and a specificity of 0.88 with leave-one-out cross-validation [212]. Hong et al. have also applied CNN for three-class classification for BE and neoplasia using endomicroscopic images with an accuracy of 80.77% [213].

However, the CAD system's performance and generalization capability are still largely constrained by the scarcity of annotated images. Besides collecting more annotated data, researchers incorporate unsupervised, semi-supervised, and weakly supervised learning to make the most of the data available.

One major application of semi-supervised learning is image segmentation. Papandreou et al. are among the first in studying weakly- and semi-supervised learning for semantic image segmentation [214]. The authors designed an expectation-maximization (EM) algorithm for iteratively updating the prediction for pixel-level annotations and the parameters of segmentation neural networks. Jia et al. approached the weakly-supervised segmentation with multiple instance learning by aggregating the pixel-level annotations and designed a constrained optimization process when additional supervision information is available [215]. More recently, Li et al. studied weakly-supervised segmentation for the prostate cancer pathological images by utilizing the prior knowledge about the epithelium-stroma distribution [216].

Besides image segmentation, semi-supervised learning has also been applied for image classification. For example, our prior work applies semi-supervised learning to eCLE images by propagating labels to unlabeled images [195]. However, handcrafted feature extraction is one major bottleneck for further performance improvement. The CAE's unsupervised nature makes it a popular choice for automated image feature extraction [217], which can naturally utilize the unlabeled data for improved feature representation. Thus, in this study, we adopt the CAE to build a semi-supervised deep neural network called CAESNet for improving the classification of BE status using OE images.

4.3 Methods

4.3.1 Data

The data we used were images collected from patients undergoing endoscope-based Confocal Laser Endomicroscopy (eCLE) procedures for BE at Emory Hospital. The dataset consists of 429 images labeled as one of the nine classes by an expert gastrointestinal endomicroscopist and 2,826 unlabeled images for semi-supervised learning. The statistics of these images are summarized in Table 4.2. Example images for each class are shown in Figure 4.2.

Table 4.2: Statistics of the BE Dataset

	Sub-Classes	Number of Images
Low Risk	Squamous (Sq)	41
	Intestinal Metaplasia (IntMet)	153
	Low Grade Dysplasia (LowG)	23
High Risk	High Grade Dysplasia (HighG)	4
	Intraepithelial Carcinoma (IntraCar)	43
Other	Duodenum (Duod)	48
	Gastric Antrum (GasAnt)	60
	Gastric Body (GasBody)	28
	Gastric Cardia (GasCard)	29
Total		429

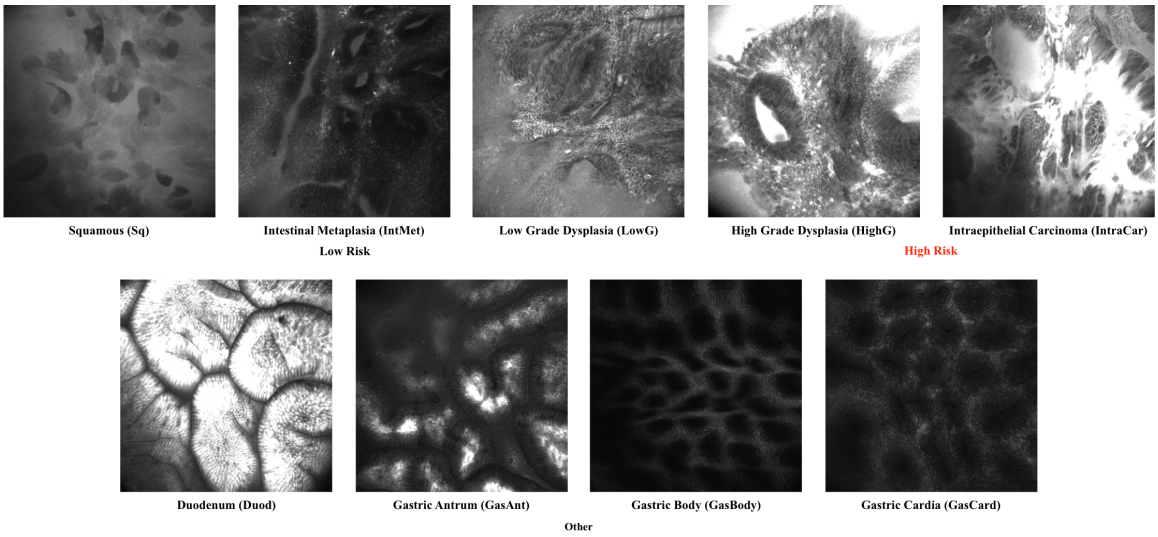


Figure 4.2: Example images of the eCLE dataset we used in this paper. The images can be classified into nine categories including squamous (Sq), Intestinal metaplasia (IneMet), Low grade dysplasia (lowG), High grade dysplasia (HighG), Intraepithelial carcinoma (IntraCar), Duodenum (Duod), Gastric antrum (GasAnt), Gastric Cardia (GasCard).

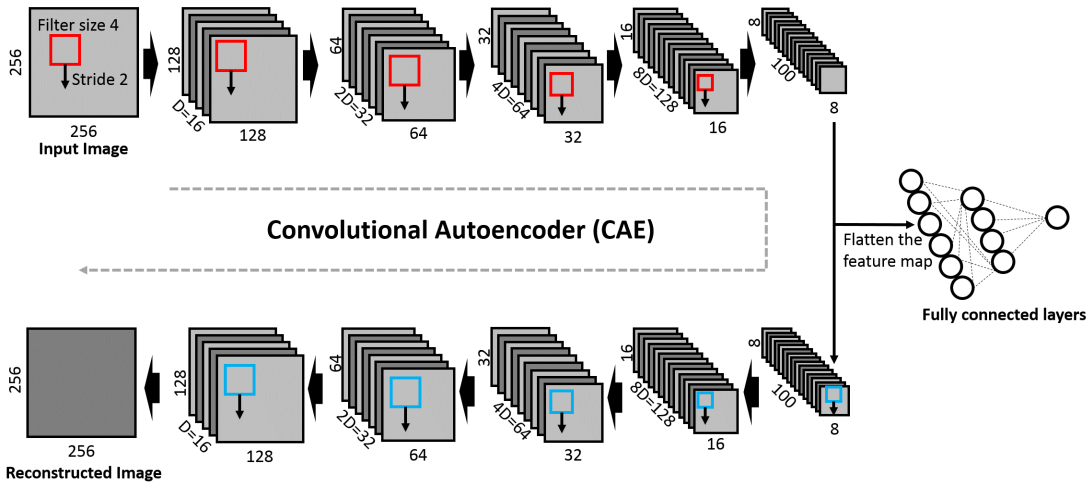


Figure 4.3: Visualization of the proposed convolutional autoencoders based semi-supervised learning model CAESNet. The original images are first encoded into hidden codes through five layers of convolutional layers with a filter size of 4×4 and a stride of 2. The hidden codes can be either feed into fully connected layers for classification or can be feed into the five layers of deconvolutional layers for decoding into original images.

4.3.2 Image Preprocessing and Data Augmentation

The original size of our eCLE images is 1024×1024. Because of the limited number of training data, we apply image augmentation to increase the number of instances for training, which is expected to improve image reconstruction and classification performance. We apply a data augmentation for each batch of training images consisting of random rotation, zooming-in, and flipping. The augmented images are then scaled into a smaller size of 256×256 to reduce the number of parameters. More details of the image preprocessing and augmentation can be found in [196].

4.3.3 CAESNet: Stacked Convolutional Autoencoders (CAEs) for Semi-supervised Learning

The structure of CAESNet is shown in Figure 4.3, which consists of a stacked convolutional autoencoder for unsupervised feature representation and fully connected layers for image classification. The encoder consists of five convolution layers, each with a filter of size 4 and stride 2, resulting in encoded hidden codes of size 8×8. Similarly, the decoder consists

of five deconvolution (transposed convolution) layers with the same filter size and stride as that of encoders. The depth of each layer of the encoder is 16, 32, 64, 128, and 100, respectively, with the last layer as the bottleneck layer. The bottleneck layer with dimension $8\times8\times100$ is first flattened into a vector of length 6,400 and then connected to the second fully connected layers through a ReLU activation function, a dropout layer, and a batch normalization layer. Finally, the label is predicted from the second fully connected layer with a soft-max function. The dropout layer is applied during the training stage and turned off during the test stage. There are two loss functions in this network, namely reconstruction loss and classification loss. We use the reconstruction loss in the unsupervised stage to train the encoder-decoder with both labeled and unlabeled images. In the supervised stage, we use the classification loss to train the encoder-classifier with only labeled images. The reconstruction loss l_R is a measure of the differences between the input images and the reconstructed images, measured by the mean squared errors:

$$l_R = \frac{1}{M} \cdot \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} (x_{ij} - \hat{x}_{ij})^2 \tag{4.1}$$

, where M is the number of images in the batch, N is the number of pixels of each input image, x_{ij} is the original value of the jth pixel of the ith image, and $\hat{x}_{ij}$ is the value of the jth pixel of the ith reconstructed image.

We use the cross-entropy function as classification loss l_C to measure of the differences between the real labels and the predicted labels of the images:

$$l_c = -\frac{1}{M} \sum_{i=1}^{M} (y_i \cdot \log(\hat{y}_i)) \tag{4.2}$$

, where M is the number of images in the batch, y_i is the real label of the ith image in the batch, and $\hat{y}_i$ is the predicted probability of the ith image. The y_i is a vector of one-hot coded labels, which equals 1 for the real label and 0 for other labels.

We construct four models utilizing all or parts of this network structure with or without

Table 4.3: The Major Components of Four Models

	Components	Loss Function	Train Set	Endpoints
Model 1	Encoder, Decoder	Reconstruction Loss	Labeled/ Unlabeled	Reconstruction
Model 2	Encoder, Fully Connected	Classification Loss	Labeled	Classification
Model 3	Encoder, Decoder, Fully Connected	Reconstruction Loss Classification Loss	Labeled	Classification
Model 4 (CAESNet)	Encoder, Decoder, Fully Connected	Reconstruction Loss Classification Loss	Labeled and Unlabeled	Classification

the unlabeled images. The major components of these four models are summarized in Table 4.3.

Algorithm 3: Unsupervised and supervised learning

Def unsupervisedTrain(Image):
while *unsupervisedTraining()* **do**
$D \leftarrow$ getRandomMiniBatch()
$z_i = Encoder_\theta(x_i)\ \forall x_i \in D$
$\hat{x}_i = Decoder_\varphi(z_i)\ \forall z_i$
$l_R = \frac{1}{M} \cdot \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} (x_{ij} - \hat{x}_{ij})^2$
$L_R = \sum_n l_R(x_i)$
$(g_\theta, g_\varphi) \leftarrow (\frac{\partial L_R}{\partial \theta}, \frac{\partial L_R}{\partial \varphi})$
$(\theta, \varphi) \leftarrow (\theta, \varphi) + \Gamma(g_\theta, g_\varphi)$

Def supervisedTrain(Image, Label):
while *supervisedTraining()* **do**
$D \leftarrow$ getLabeledRandomMiniBatch()
$z_i = Encoder_\theta(x_i)\ \forall x_i \in D$
$\hat{y}_i = FC_\phi(z_i)\ \forall z_i$
$l_C = -\frac{1}{M} \sum_{i=1}^{M} (y_i \cdot \log(\hat{y}_i))$
$L_C = \sum_n l_C(x_i, y_i)\ \forall \{x_i, y_i\} \in D$
$(g_\theta, g_\phi) \leftarrow (\frac{\partial L_C}{\partial \theta}, \frac{\partial L_C}{\partial \phi})$
$(\theta, \phi) \in (\theta, \phi) + \Gamma(g_\theta, g_\phi)$

In model 1, we implemented a stacked CAE, which is an unsupervised feature representation network. Each input image is encoded by an encoder with five convolution layers into hidden codes. The input images are then reconstructed by a decoder with five deconvolution layers (transposed convolution layers) from the hidden codes. The network is trained

Algorithm 4: Training schemes for four models

```
Def Model_1(UnlabeledImages):
while modelCoverged() do
    unsupervisedTrain(UnlabeledImages)
```

```
Def Model_2({LabeledImages, Labels}):
while modelConverged() do
    supervisedTrain({LabeledImages, Labels})
```

```
Def Model_3({LabeledImages, Labels}):
while modelConverged() do
    unsupervisedTrain(LabeledImages)
    supervisedTrain({LabeledImages, Labels})
```

```
Def Model_4(UnlabeledImages, {LabeledImages, Labels}):
while modelConverged() do
    unsupervisedTrain(UnlabeledImages + LabeledImages)
    if LabeledImages then
        supervisedTrain({LabeledImages, Labels})
```

by minimizing reconstruction loss (Algorithm 3 unsupervisedTrain).

In model 2, we use only the stacked convolutional encoders and the fully connected layers for image classification, which is a standard implementation for image classification. Each input image is firstly encoded and then classified into multiple categories. The network is trained by minimizing the classification loss using only the labeled images (Algorithm 3 supervisedTrain).

In model 3, we use both the stacked CAE for image reconstruction and the fully connected layers for image classification, but only using the labeled images. In each training step, we first update the encoder and decoder by reconstruction loss, then we update the encoder and fully connected layers by classification loss. All images are firstly encoded into hidden codes. After encoding, we reconstruct the input images and update the encoder and decoder by taking the gradient of reconstruction loss. Then, we use the updated encoder to regenerate the hidden codes and pass them through fully connected layers for image classification. We update the encoder and fully connected layers by taking the gradient of the classification loss (Algorithm 4 Model_3).

In model 4 (CAESNet), we use the same structure as model 3 but utilize both labeled and unlabeled images for training. For each image, if it is labeled, we first update the encoder and decoder, then update the encoder and fully connected layers; on the other hand, if it is unlabeled, we only update the encoder and decoder by minimizing the reconstruction loss (Algorithm 4 Model_4). The implementation details of model 4 are presented in Supplementary Note 2.

4.3.4 Model Evaluation and Classification Metrics

The three classification pipelines are evaluated using stratified four-fold cross-validation. We split the labeled data into training, validation, and test datasets in each fold of the cross-validation. We initiate the weights of our networks with multiple random seeds and select the best model based on the classification loss on the validation set. Data augmentation has been applied only on the training datasets. To enable statistical test, we repeat the four-fold cross-validation three times.

We evaluate the multi-class classification results using four metrics, namely accuracy, precision, F1 score, and Cohen Kappa score. We do not include recall since it is the same as accuracy for multi-class classification if we weighted the recall by the number of samples in each class.

4.3.5 Experiment Configuration

We run all data processing pipelines and models on a single server with multiple CPU cores and two NVIDIA Tesla K80 GPUs. The image augmentation is implemented with Python Library Augmentor [218] with random rotation, scaling, and flipping. The CAESNet and the baseline deep learning models are implemented with PyTorch. The classification results are evaluated with scikit-learn [179]. For the training of semi-supervised networks, we use a batch size of 20, epochs of 200, and a learning rate of 0.0002. We apply different depths (i.e., 16, 32, and 64) to each convolutional layer to see the influence on the classification

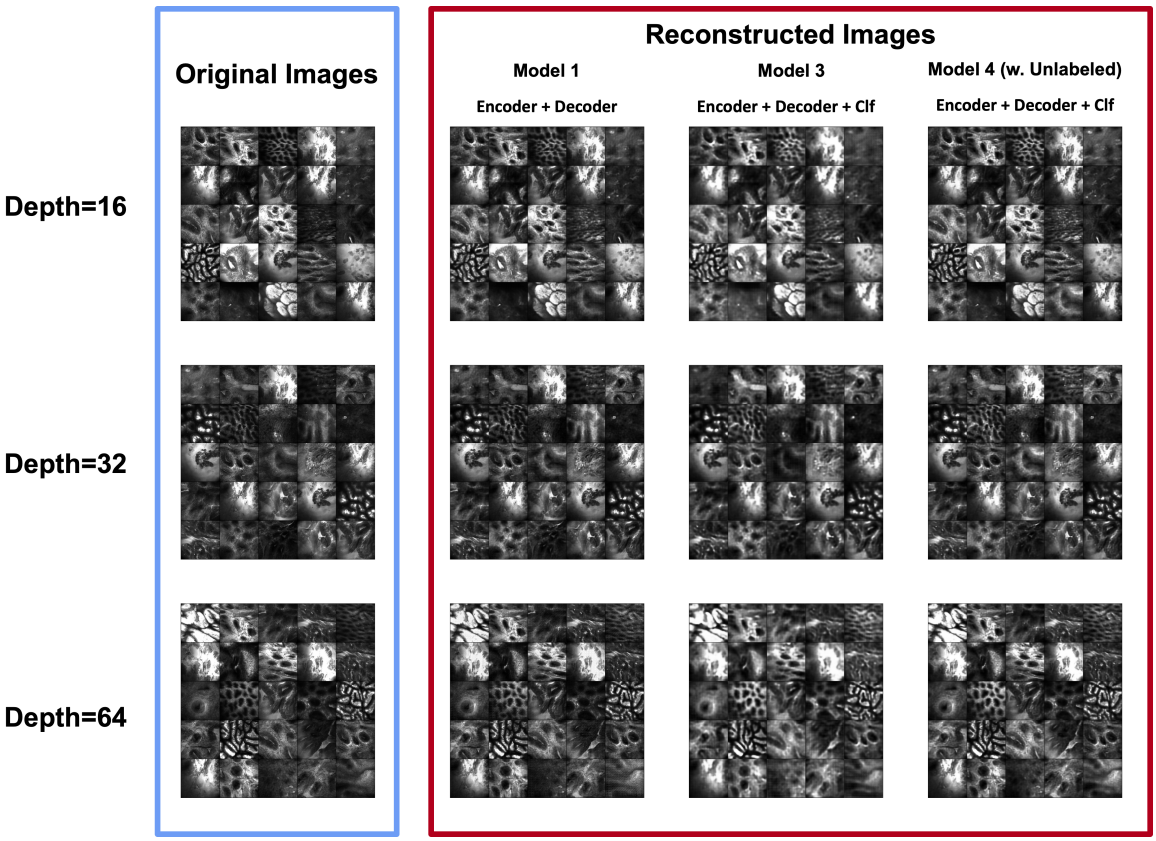


Figure 4.4: The original images (blue rectangle) and the corresponding reconstructed images (red rectangle) by autoencoders in three different models. Model 1: an autoencoder using labeled images. Model 3: an autoencoder + a classifier using only labeled images. Model 4 (CAESNet): an autoencoder + a classifier using both labeled and unlabeled images. Model 1 and model 4 (CAESNet) achieve better reconstruction results compared to those of model 3.

and reconstruction performance.

4.4 Results

4.4.1 Improved Image Reconstruction Performance by Extra Training with Unlabeled Images

The image reconstruction quality increases when increasing the number of training images. Figure 4.4 shows the original images and the reconstructed images from the test set for models 1, 3, and 4. The reconstructed image quality of model 3 is the worst compared to model 1 and model 4. However, model 4 can achieve similar reconstruction performance

Table 4.4: Classification Performance of Models with Various Implementations

Category	Imp.	Accuracy	Precision	F1	Cohen Kappa
Graph-Based (Baseline)	LP	0.734±0.0175	0.763±0.0479	0.674±0.0269	0.642±0.0241
	Rand. LP(20)	0.652±0.0465	0.709±0.0277	0.584±0.0453	0.528±0.0741
	Rand. LP(90)	0.686±0.0318	0.702±0.0519	0.624±0.0394	0.577±0.0477
Model 2 (Supervised)	Depth 16	0.679±0.0375	0.67±0.0623	0.652±0.046	0.6±0.0491
	Depth 32	0.683±0.0368	0.671±0.0661	0.656±0.0458	0.6±0.0486
	Depth 64	0.667±0.0431	0.628±0.0784	0.626±0.0554	0.572±0.0601
Model 3 (Supervised)	Depth 16	0.787±0.0369	0.804±0.0376	0.774±0.0354	0.733±0.0439
	Depth 32	0.808±0.045	0.819±0.034	0.798±0.0448	0.761±0.052
	Depth 64	0.816±0.0384	0.822±0.0446	0.807±0.0425	0.768±0.0498
Model 4 (Semi-Supervised)	Depth 16	0.784±0.046	0.789±0.0419	0.776±0.0458	0.731±0.0542
	Depth 32	**0.824±0.0329**	**0.832±0.0302**	**0.816±0.0342**	**0.781±0.04**
	Depth 64	0.815±0.0252	0.814±0.0265	0.804±0.0262	0.768±0.0306

as model 1, even with an extra classification task. The poor reconstruction performance of model 3 may result from the trade-off between image reconstruction and classification, and the limited number of labeled images.

4.4.2 Improved Classification Performance by Semi-Supervised Learning with Unlabeled Images

The classification performance of three models is shown in Figure 4.5 and Table 4.4. Based on Table 4.4, the performance of model 2 is inferior compared to the best baseline model (LP). However, model 3 and model 4 consistently achieve better performance than the baseline models at all network depths, likely resulting from utilizing the reconstruction loss for regularizing the network. Thus, when trained with the same amount of labeled data, only model 2 suffers from underfitting. Model 3 and model 4 achieve similar prediction performance, where model 4 at depth 32 achieves the best average performance ($0.824/pm0.0329$). We have also performed the pair-wise two-sample t-test for all models' prediction performance in Table 4.4. Model 3 and model 4 significantly outperform model 2. However, no significant difference has been identified between model 3 and model 4. In Figure 4.5, we visualized the significance levels between models 2, 3, and 4 at the same network depth.

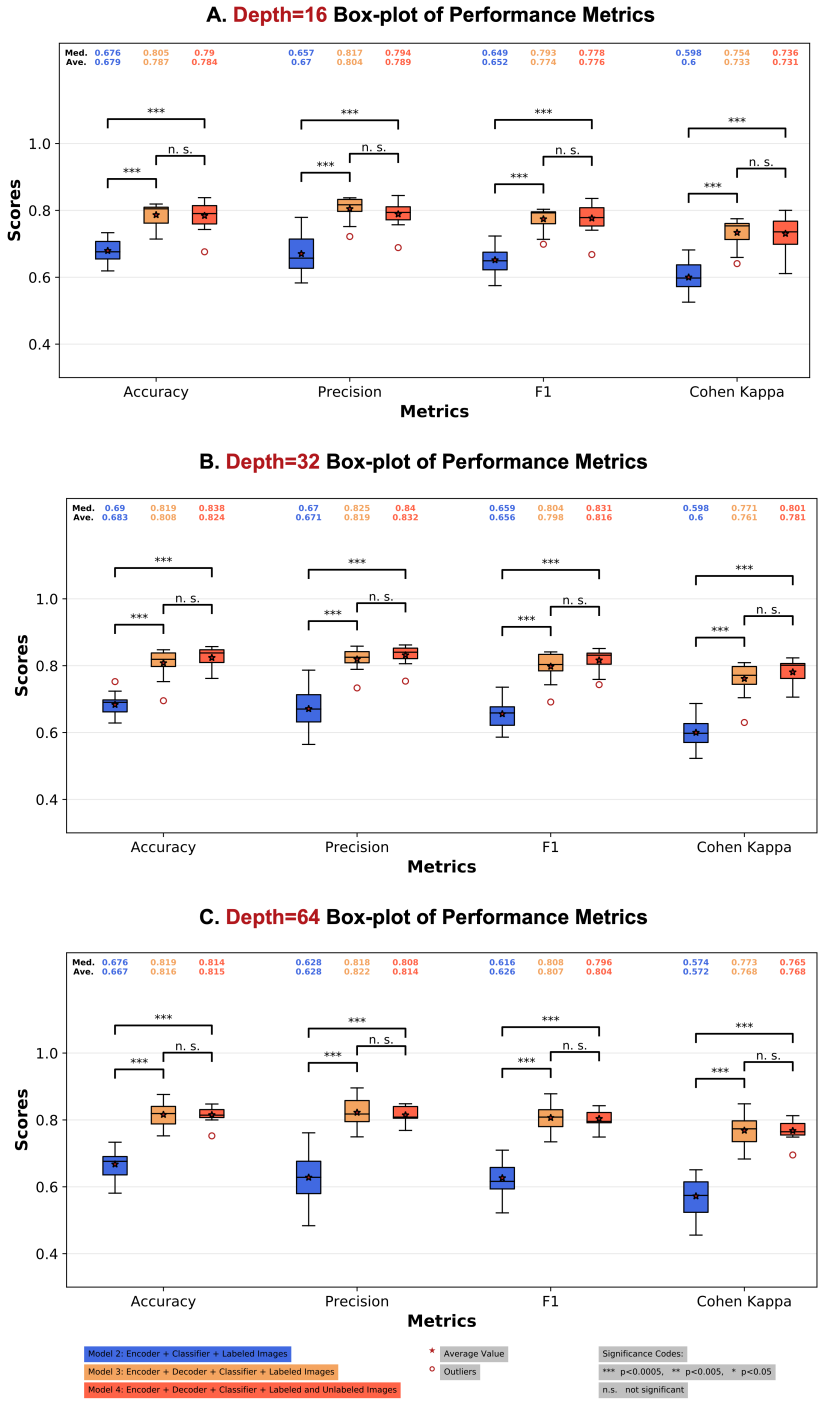


Figure 4.5: Boxplot of the classification performance of various models and depths. A. The classification performance of three models using depth 16; B. The performance of three models using depth 32; C. The performance of three models using depth of 64. Model 3 and model 4 (CAESNet) achieve similar prediction performance and consistently outperforms model 2. Model 4 (CAESNet) at depth 32 achieves the best average performance (0.824 ± 0.0329).

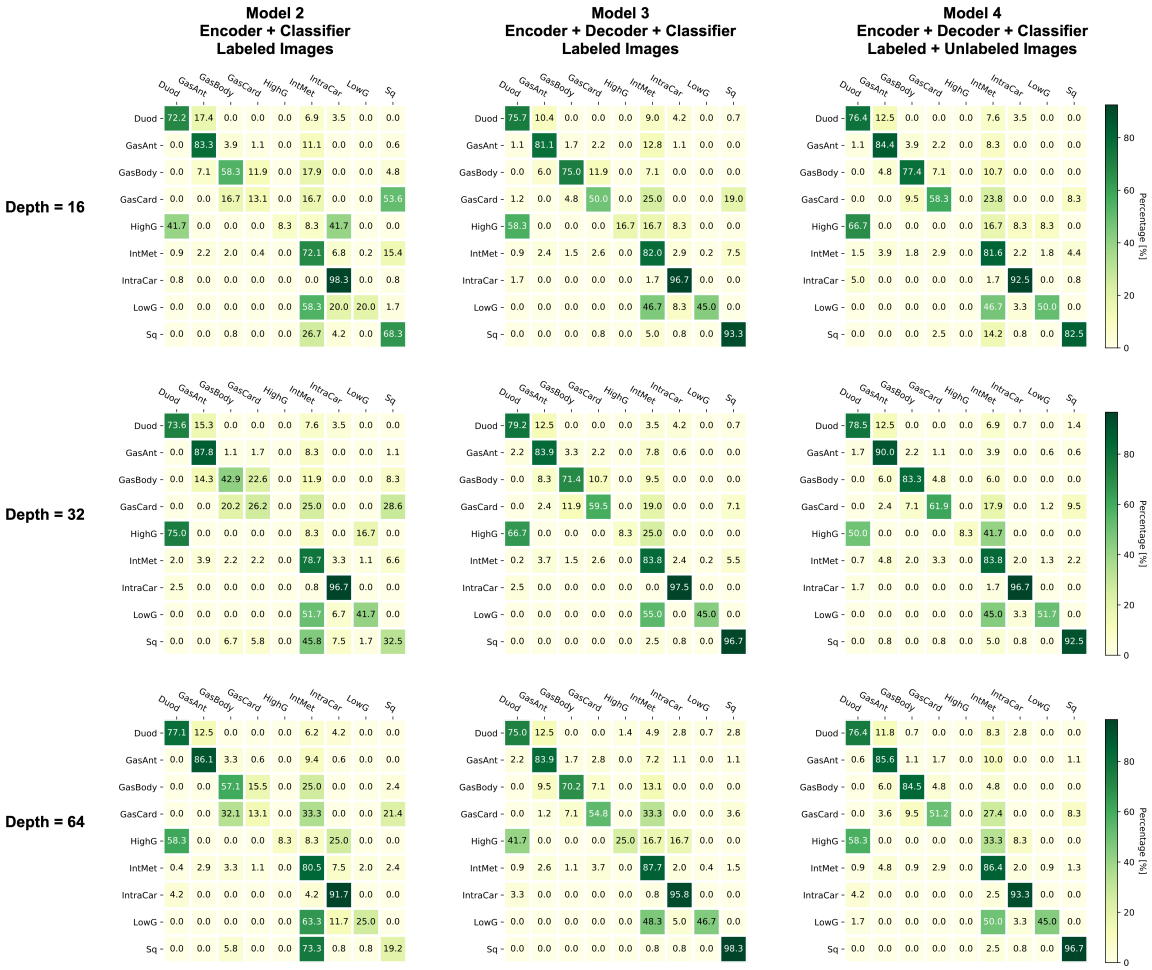


Figure 4.6: Confusion matrices of various models and depths. Each confusion matrix is color-coded as a heatmap for visualization purpose. The model 3 and model 4 (CAESNet) consistently achieves better performance compared to the model 2 at all network depths.

The confusion matrices of three models with various depths are shown in Figure 4.6. An ideal classification should achieve the diagonal pattern in the confusion matrix. Model 3 and model 4 concentrate more on the diagonal cells than model 2, which is consistent with their overall performance. However, both model 3 and model 4 makes a relatively poor classification for Gastric Cardia (GasCard), Low-Grade Dysplasia (LowG), and High-Grade Dysplasia (HighG). These three classes are tended to be misclassified as Intestinal Metaplasia (IntMet). We suspect the limited numbers of samples could cause this misclassification for GasCard, LowG, and HighG, where we only have 29, 23, and 4 samples out of the 429 labeled images, respectively. On the contrary, we have 153 IntMet samples, which might dominate the classification decisions.

4.4.3 Fluctuated Performance with the Number of Unlabeled Images

After confirming that utilizing unlabeled images to help the regularization of the encoder can improve the prediction performance, we want to investigate the influence of the number of unlabeled images utilized for the model training. With the 2,826 unlabeled images, we feed the model in an accumulative fashion. For the ratio increasing from 0 to 1.0 with a step size of 0.1, we always use the first proportions of the unlabeled images. For example, when the ratio equals 0.1, we use the first 10% of unlabeled images; when the ratio equals 0.2, we use the same first 10% of unlabeled images plus the subsequent 10% of unlabeled images. There are fluctuations in the prediction performance along with the number of unlabeled images applied. We have also investigated the influence of data augmentation on classification performance using model 2, where similar performance fluctuations have been observed.

4.5 Conclusion and discussion

In this study, we developed CAESNet, a CAE-based semi-supervised learning framework for the multi-class classification of endomicroscopic images. We conclude that the stacked

CAE is an effective deep learning method to extract informative features from the eCLE images based on the extensive experiments. The CAE network allows us to add a regularization for the classification loss and allows us to utilize the unlabeled images to optimize the encoder in a semi-supervised learning fashion.

When utilizing a different number of unlabeled images, the performance does not follow a monotonic increasing pattern but fluctuates as the number of unlabeled images increased. There are multiple potential explanations. First, the results could be caused by the training configurations where we apply the same hyperparameters for experiments with different numbers of unlabeled images. To solve this issue, we need to search optimized configurations for each model and dataset. Second, the performance may be related to the quality and underlined labels of the unlabeled images. If the unlabeled images are less relevant to the labeled images, we may experience an adverse effect when utilizing these unrelated unlabeled images. We also suspect that if most unlabeled images are from a specific class, they may preoccupy the autoencoder and make it overfit images from that class. One possible way to solve this problem is to introduce another coefficient to balance reconstruction loss and classification loss. When we have a larger number of unlabeled images, the reconstruction loss typically drops much faster than classification loss. Thus, we can assign larger weights to classification loss so that we can balance the training of image classification and image reconstruction.

There are multiple future directions for our semi-supervised model. One future direction is to improve the unsupervised feature representation by applying adversarial autoencoders (AAE). AAE is a probabilistic autoencoder by matching the aggregated posterior of the hidden code vector of the autoencoder with an arbitrary prior distribution [219]. As a result, the encoder tends to generate more meaningful hidden codes. Currently, we only use the unlabeled data to improve the unsupervised feature representation with simple handcrafted data augmentation. With the AAE framework, we can additionally auto-augment the labeled images from a few labeled images and the unlabeled images [220].

In this method, we learn a generator for sequences of incremental, black-box transformation functions (TFs) from the unlabeled images and then apply the learned TF generator to the labeled images for augmentation of realistic labeled images. This advanced auto-augmentation for labeled images should be able to improve the classification performance.

Another direction is introducing interpretation for the deep model to enable clinical translation. By introducing techniques like the attention to visualize the model, we want to identify the features that contribute most to the correct classification. This process can, in turn, serve as a parameter tuning or diagnosis step for the model. By examining whether the neural network is truly picking up the disease-relevant structures for the prediction, we can differentiate models fitting the data noise from the truly effective models.

CHAPTER 5
WEAKLY-SUPERVISED LEARNING FOR HISTOPATHOLOGICAL IMAGING INFORMATICS

5.1 Introduction

Whole-slide imaging (WSI) generates digital slides by scanning conventional glass slides, which enables the computer-aided diagnosis (CAD) for pathological images and the integration of digital pathology with high-dimensional genomics features [221]. The virtual slide with a resolution almost at the optical resolution limits can be generated in sixty seconds for an entire glass slide [221]. The resulting virtual slides consist of digital images of histological sections over a complete range of standard magnifications. Extensive validations by the College of American Pathologists Pathology have demonstrated WSI's effectiveness for diagnostic interpretation [222]. Manual examination of slides and diagnosis by pathologists can be subjective and time-consuming because of the large fields-of-view to be reviewed. With the adoption of WSI systems in clinical settings, researchers have built numerous CAD systems for various clinical endpoints. The development of WSI and the corresponding computational algorithms forms two major components of digital pathology, which has paved the way towards precision medicine. Due to the large size of image data contained in a virtual slide, it is impractical to process the entire digital image of a slide as a whole. Most CAD systems first select the region of interest (ROI) using low-resolution thumbnails and then tile the image within ROIs into small image patches at the highest magnification level. After the tiling process, the image patches are processed with feature extraction, feature selection, and predictive modeling for conventional digital image processing pipelines. With the huge success of deep learning in natural images, models like deep convolutional networks (ConvNets) have also been applied to these image patches for

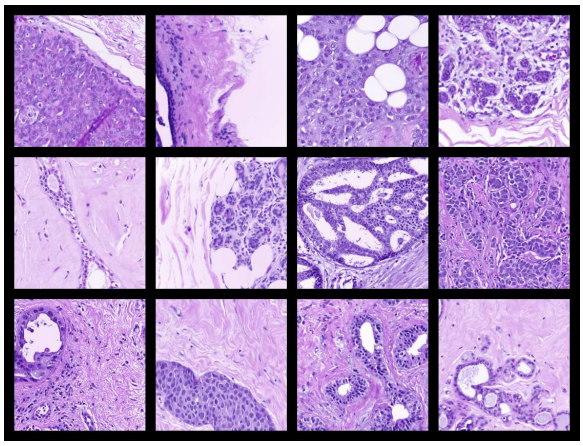
efficient feature representation. The features generated by ConvNets are then fed into extra networks for segmentation and classification. Recent works have demonstrated promising results for applying deep learning to WSI [40, 42].

However, current CAD models typically process the image patches independently without considering each patch's geographical relationships. This paradigm results in two major drawbacks. First, the image patches are processed without the support from neighboring image patches, which may contain useful context information. For example, specific structures may be partially cropped in a specific image patch and reduce the power of feature representation. Second, only the finest images at the highest magnification are used for feature extraction, while the coarser images at lower magnifications are discarded completely. These coarser images also carry useful information and should have been utilized for the feature representation. To address these drawbacks of current CAD models for pathological images, we propose exploiting each image patch's context information with multi-scale ConvNets. Besides the basic image patch tiled from the finest resolution, we also extract image patches from two lower resolutions at the same geographic center to provide context information (Fig. 5.1). These two extra image patches are captured from the WSI image pyramid's coarser levels and thus have larger Field of Views (FOVs) compared with the original image patches. We build three ConvNets for three concentrical image patches and then combine the extracted features for the final classification of the primary image patch. We have applied our multi-scale ConvNets to a benchmark breast cancer dataset. Extensive experiments have demonstrated that our multi-scale ConvNets can significantly improve the classification performance of breast cancer.

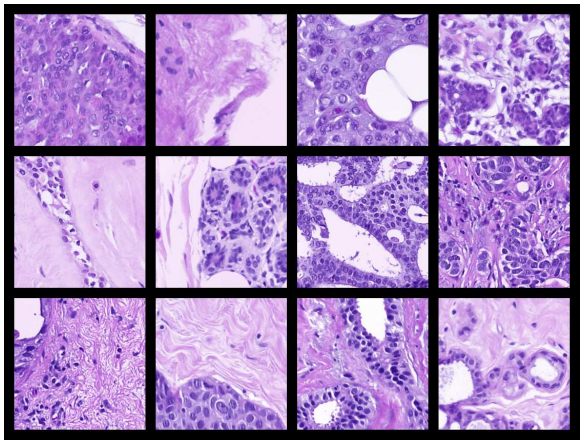
5.2 Related Works

5.2.1 Computer-aided diagnosis for WSI

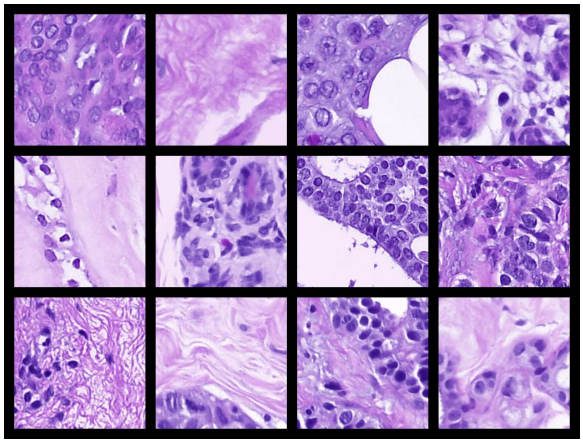
The development of CAD systems for digital pathology has become one of the fastest-growing research fields due to WSI's growing popularity. A typical CAD pipeline for WSI



(a) FOV 1024



(b) FOV 512



(c) FOV 256

Figure 5.1: Visualization of concentric images with different FOVs. From (a) to (c), the FOV is getting smaller and smaller, with more zoomed in view of the same image patch.

consists of four major components: 1) image quality control, 2) feature extraction at the pixel, object, and semantic levels, 3) predictive modeling using imaging features, and 4) model visualization for interactive discovery [35].

Image quality control is an essential step for digital pathology because of the heterogeneities of WSIs collected between different clinical sites with various platforms and slide preparation protocols, which is the so-called batch effect. On the other hand, the WSI may also have artifacts, including tissue folds, blurred regions, pen marks, and shadows. The batch effects and image artifacts have unpredictable effects on image segmentation, classification, and other quantitative image analysis tasks. Researchers have developed multiple techniques, including color normalization, scale normalization, and blur detection, to eliminate or correct the batch-effect and image artifacts of WSIs.

Feature extraction is another essential step to represent the WSI data quantitatively. Conventional digital imaging processing techniques extract features from pathological images at pixel and object levels to capture the morphological properties [36]. Pixel-level feature extraction identifies the properties of color and texture for all image pixels. Color features are typically expressed with the color spread, prominence, and co-occurrence using statistics and frequencies of color histograms in different color spaces. Texture features quantify image sharpness, contrast, changes in intensity, and discontinuities or edges by measuring properties from gray-level intensity profiles. Object-level feature extraction requires the segmentation of cellular structures and captures the shape, texture, and spatial distribution of cellular structures in a WSI. Besides the features extracted from WSI, researchers also proposed to integrate the pathological features with clinical features and genomic features for improved diagnosis [37].

However, the conventional feature extraction relies heavily on hand-crafted features, limiting the generalizability of the features. With deep learning development, the human-designed feature extraction has been replaced by feature representation with deep neural networks. For imaging data, the most popular feature representation network is the convo-

lutional neural network (ConvNets). Deep ConvNets can learn efficient feature representation from a large amount of training data. By combining ConvNets with fully connected (FC) layers and a softmax layer for classification, the deep networks can be trained end-to-end and thus can learn both feature representation and classification from the training data. With the success of deep learning in natural images, deep neural networks like ConvNets have been applied to medical images including MRI (brain tumor [38]), CT (lung nodule [39]), and WSI (breast cancer [40], lung cancer [41], glioma [42], heart rejection [43] etc.).

5.2.2 Image pyramids and multi-scale models for natural images

The image pyramids applied in object recognition inspire our multi-scale ConvNets for WSI. Image pyramids consist of multi-scale representations of the same image. Feature pyramids built upon image pyramids are among the most standard solutions for object recognition at various scales in computer vision. The object's scale change is offset by its level within the pyramid so that the objects can be recognized in a scale-invariant fashion. By scanning a trained model over both positions and pyramid levels, objects across a large range of scales can be robustly detected. Before the deep learning era, dense sampling on image pyramids and hand-crafted features are critical for accurate object detection. With the success of deep learning for natural images, the hand-crafted feature extraction step has been replaced by the powerful feature representation using deep ConvNets. Although ConvNets are much more robust to scale variances than hand-crafted features, image pyramids are utilized to ensure the most accurate performance. For example, multiple top entries in the ImageNet [223] and COCO [224] detection challenges exploit multi-scale inference with image pyramids.

Besides the direct application of image pyramids for multi-scale inference, recent works have also utilized the built-in feature hierarchy of deep ConvNets for multi-scale feature representations. For example, the feature pyramid network (FPN) uses the pyramid shape of a ConvNet's feature hierarchy. It builds semantically strong multi-scale representations

with the bottom-up pathway, top-down pathway, and lateral connections [225]. The FPN gets rid of image pyramids and creates in-network feature pyramids, which significantly reduces the speed and memory without sacrificing the multi-scale representation power. Multiple works, including Mask R-CNN [226], has utilized the FPN framework to achieve improved performance.

One major difference between our work and the multi-scale object detection is that we aim to use the built-in image pyramids of WSI and improve the prediction performance with the support of context information. Thus, we improve the current model by enlarging the FOVs using images of coarser levels on the WSI image pyramids.

5.2.3 Multi-scale features for medical images

Xu et al. have applied multi-scale context-aware networks for colorectal liver metastases [227]. They utilized the context information from low magnification levels by concatenating the feature maps generated by deep convolutional neural networks at early and late stages. The combined features improved the performance for image segmentation and classification tasks.

5.3 Multi-scale convolutional networks

5.3.1 Late Fusion

One intuitive way to integrate images from different scales is to combine the hidden feature vectors extracted from the three concentric images. Since the integration happens before the last FC layer, we call this family of methods as of late fusion (Fig. 5.2). We first use the five layers of ConvNets and two FC layers to extract features from input images, respectively. After the five layers of convolution, each image is represented with a feature map with size $256 \times 6 \times 6$. We then flatten the feature maps and get a vector with a length of $4,096$ for each image after the two FC layers. Finally, we combine the feature vectors before feeding the last classification layer by concatenating or taking the average. The

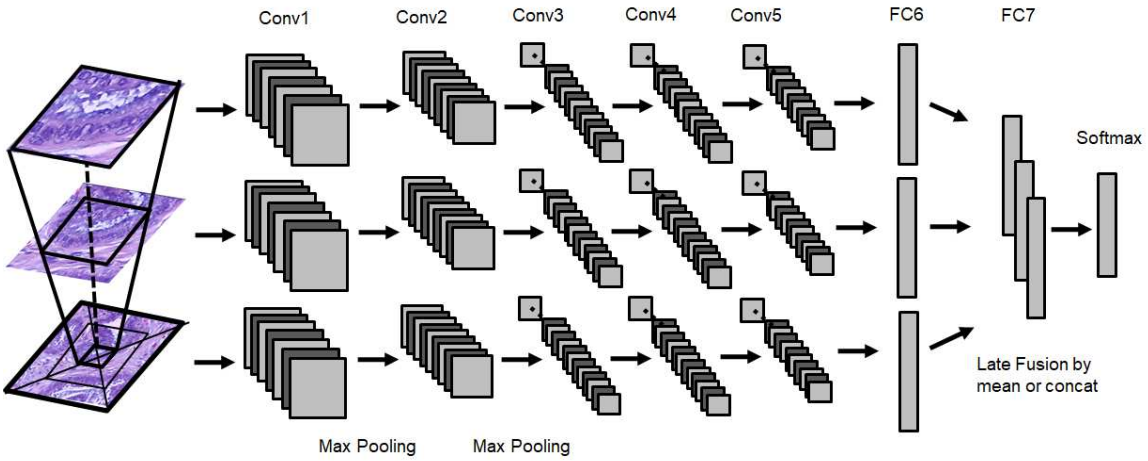


Figure 5.2: Visualization of multi-scale patch integration by late fusion. The information extracted from each scale is not combined until feeding into classification layer. The feature vectors represented from each scale of concentric image patches are integrated by either concatenation or taking average. The combined feature vectors are then fed into last FC layer for classification.

combined feature vectors then go through another FC layer for classification.

5.3.2 Early Fusion

Another way to integrate these multi-scale images is to combine the feature maps at a relatively early stage within the five layers of ConvNets. We call these methods early fusion. Based on the methods we use to fuse the feature maps, we can further classify them into full concatenation and partial fusion.

Full concatenation directly concatenates the feature maps of different concentric images (Fig. 5.3). Since the images are scaled to the same dimension ($3 \times 224 \times 224$) and parsed through ConvNets of the same structure, the feature maps also have the same dimension. We can directly concatenate these feature maps along the depth dimension and then feed the combined feature maps to the rest of the networks. Since the concatenation will increase the depth by three times, the following ConvnNet is modified correspondingly. To simplify the model, we share all network parameters for three levels of images.

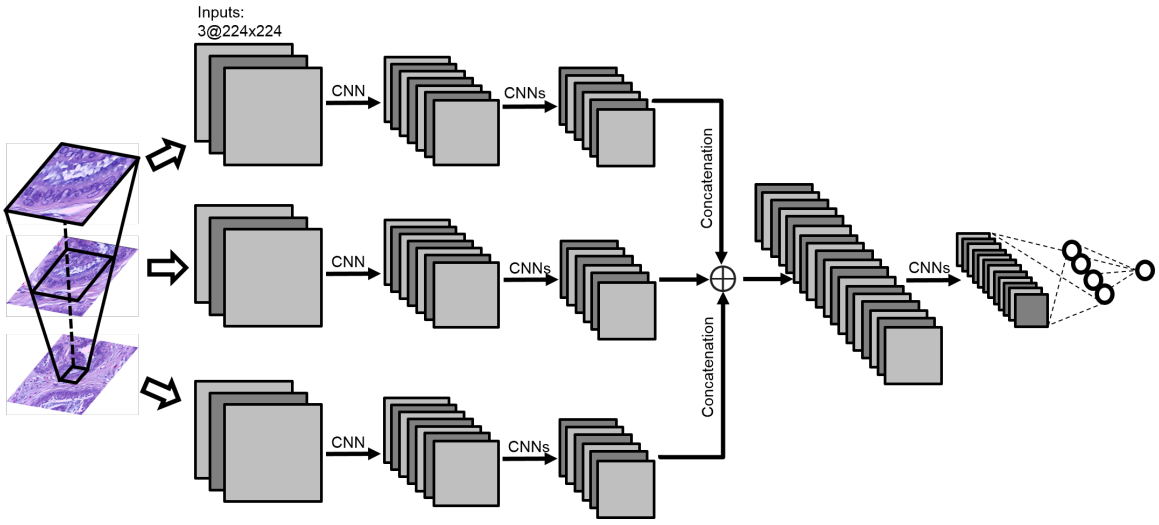


Figure 5.3: Visualization of multi-scale patch integration by early fusion and full concatenation. The feature maps generated by convolutional neural networks are fully concatenated. We have tried to concatenate the feature maps at the third and fourth ConvNet layers respectively.

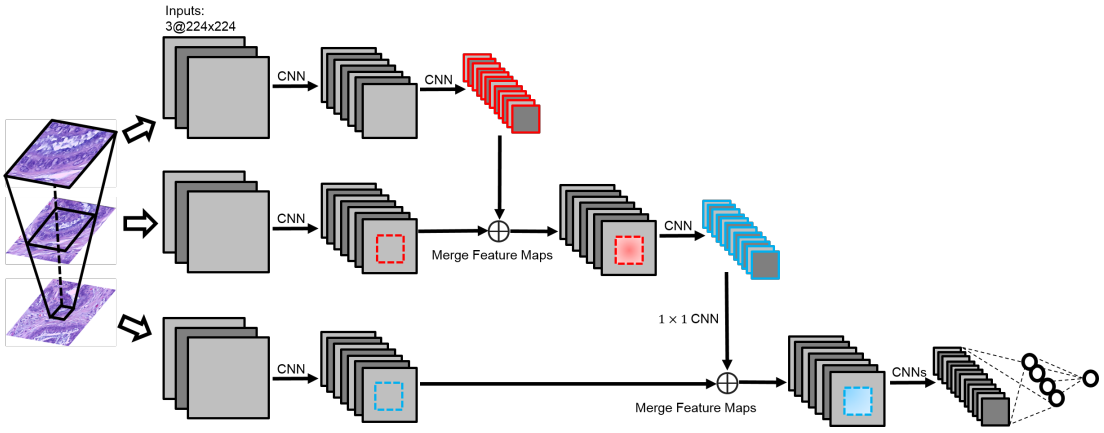
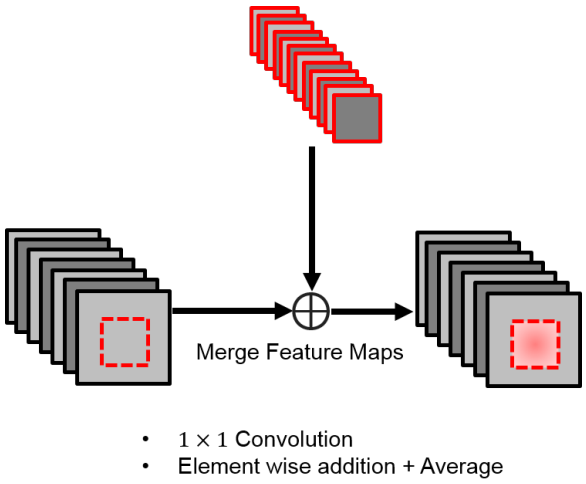
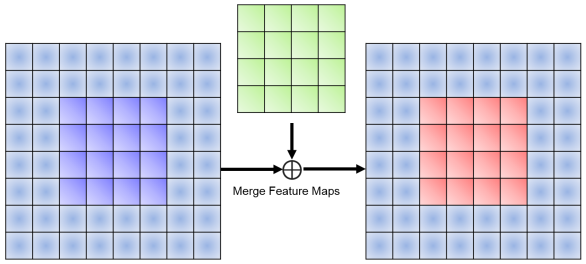


Figure 5.4: Visualization of multi-scale patch integration early partial fusion. We fuse the feature maps of multi-scale images sequentially. The image with smaller field of view is processed with two ConvNet layers and then fused with the image processed with one ConvNet layer.



(a) Partially merge the feature maps by 1×1 convolution and element wise average.



(b) The illustration of partially overlapped feature maps based on field of views.

Figure 5.5: Details of partial fusion methods.

Table 5.1: Experiment Results for Multi-Scale ConvNets

Multi-Scale	FOV-256	FOV-512	FOV-1024	Fusion	ImageNet-Pretrained	Accuracy
Single Scale	+	-	-	-	-	35.50 ± 4.04 %
	+	-	-	-	+	$68.75 \pm 3.20\%$
	-	+	-	-	-	$29.00 \pm 7.35\%$
	-	+	-	-	+	$78.75 \pm 2.99\%$
	-	-	+	-	-	$30.75 \pm 9.60\%$
	-	-	+	-	+	$74.50 \pm 5.57\%$
Late Fusion	+	+	+	Z concat	+	$81.50 \pm 2.38\%$
	+	+	+	Z mean	+	$79.75 \pm 1.89\%$
Early Fusion	+	+	+	Conv3 concat	+	$80.50 \pm 3.41\%$
	+	+	+	Conv4 concat	+	$79.25 \pm 4.27\%$
	+	+	+	Partial fusion	+	$72.00 \pm 4.08\%$

Partial fusion aims to consider the various FOVs when integrate the multi-scale feature maps (Fig. 5.4). The feature maps are merged sequentially. One image with smaller FOVs is firstly passed through two ConvNets to get smaller feature maps. While the other image with larger FOVs is passed through one ConvNets and get larger feature maps. The depth of smaller feature maps is reduced to match that of larger feature maps by 1×1 convolution (Fig. 5.5a). Then they are partially merged by aligning them at the center and take means of the overlapped elements (Fig. 5.5b). After two partial fusions to combine the three multi-scale images, the integrated feature maps are processed by the remaining three layers of ConvNets and three FC layers for classification.

5.4 Experiments

5.4.1 Datasets

The dataset we use in this study is Part A - microscopy images from ICIAR 2018 Grand Challenge on Breast Cancer Histology images [229]. The dataset contains 400 microscopy images, labeled as one of the following classes: normal, benign, *in situ* carcinoma and invasive carcinoma. Each class contains 100 images. The size of each image is 2048×1536 pixels. Two medical experts performed the labeling of images, and label disagreements

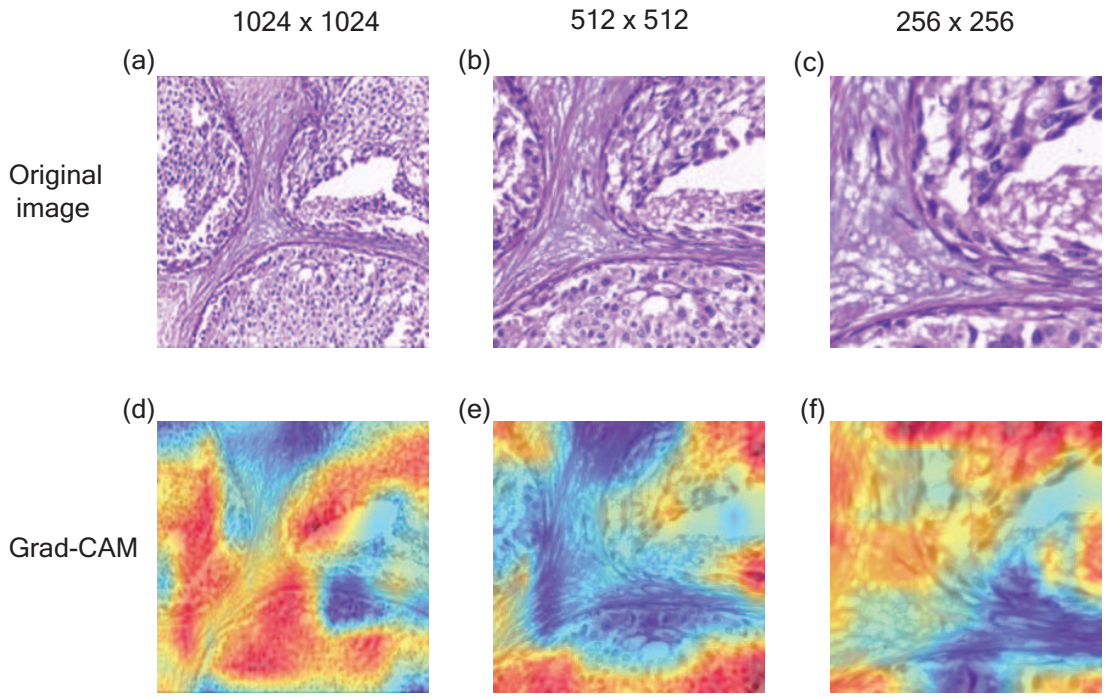


Figure 5.6: Visualization of important regions for predictions in different scales of an example image using Grad-CAM [228].

have been excluded from the dataset.

5.4.2 Data Normalization

Because of the variations in preparing each tissue sample and scanning for microscopic examination, we found significant visual differences in the histology images. To prevent this artifact from affecting our classification results, we use a color normalization method based on non-negative matrix factorization (NMF) [230]. Specifically, the method calculates stain color bases and stain densities in an unsupervised manner for a given image. We rebase the color distribution of each image by multiplying reference color bases and normalized stain densities. Our study uses Image b027 as the reference image.

5.4.3 Data Augmentation

Given that we have only limited image samples compared to typical DL applications, we augment our data to increase variations of the data and to reduce over-fitting when training our models. Specifically, we rotate the original image and crop it to a desired size (either 256×256, 512×512, or 1024×1024 pixels) (Fig. 5.1). We make sure that each cropped image will not contain more than 30% of the white background. We resize each cropped image to 224×224 pixels to fit the input size of AlexNet, the backbone network architecture we use.

5.4.4 Settings of hyperparameters

We use four-fold cross-validation and set 300 images as the training set and the remaining 100 images as the test set. Both the training and test sets have an even distribution of four class labels. During training, the batch size is set to 60. We use stochastic gradient descent (SGD) and set the momentum to 0.9 for optimization. We set a relatively large learning rate of 0.01 when we do not use ImageNet-pretrained parameters and a relatively small learning rate of 0.0002 when we use ImageNet-pretrained parameters. We record the best accuracy for the test set during 200 training iterations for each fold and report the mean and standard deviation best accuracy for all folds.

5.5 Results

We list the prediction accuracy in (Table 5.1). Below we talk about our major findings.

5.5.1 ImageNet Pretraining

We found out that using ImageNet-pretrained parameters to initialize our model significantly improved the prediction accuracy. In Table 5.1, the increase is 33.25%, 49.75%, and 43.75% under the FOV as 256×256, 512×512, and 1024×1024, respectively. ImageNet

is a large image database that contains 1.28 million images that belong to $1,000$ classes. The classes cover a wide variety of common objects such as hen and cat. Although general image classification is very different from pathological image classification, the large and comprehensive ImageNet data set enables the convolutional filters of AlexNet to capture the heterogeneous and basic visual components in the world. Therefore, ImageNet pre-trained parameters set the stage for training DL models for pathology image classification, given the limited number of pathology images in the Challenge.

5.5.2 Field of View

From Table 5.1, we see that choices of FOV affect prediction accuracy. Specifically, when we set a small FOV (256×256 pixels), the prediction accuracy is lowest with ImageNet pre-training among three sizes of FOV. With ImageNet pretraining, the medium FOV (512×512 pixels) achieves the best accuracy 78.75% among single-scale inputs. To better understand the effect of FOV, we visualize an example image in three scales and highlight important regions for AlexNet prediction with Grad-CAM [228]. We argue that although large FOV includes the most context information for classification, they lose the most details when we resize the input to fit the specified input size of AlexNet (224×224). On the other hand, small FOVs (256×256) are most likely to miss distinctive regions for identifying cancers. The medium FOV (512×512) maintains a good balance between context and detailed information (Fig 5.6).

5.5.3 Multi-scale Input

Table 5.1 shows that multi-scale input increases prediction accuracy up to 81.5% when we combine the features at a late stage by concatenating feature vectors before the last fully connected layers and share the parameters of AlexNet backbone for each scale of input. The difference in accuracy between the two late fusion methods we use, concatenation and averaging of the feature vector before the last fully connected layers, is minor (81.5%

and 79.75%). The difference between late fusion and the two early fusion methods by concatenating feature maps at Conv layer 3 and Conv layer 2 is also small (within 1.25%). However, partial fusion accuracy is only 72%, which is less than that of single scale input at FOV of 512 and 1024. We reason that partial fusion's inferior performance may be caused by the interruption of the original AlexNet structures. With a limited number of training data, the original AlexNet structure with pertained parameters contributes most to the classification performance.

5.6 Conclusions and Discussions

In this study, we have built multi-scale ConvNets for late fusion and early fusion of the WSI image pyramids to improve breast cancer classification. Based on the extensive ablation experiments, we found that the late fusion method by taking the average of feature vectors and sharing AlexNet pre-trained parameters reaches the highest accuracy (80%). The proposed method beats the other methods, including late fusion by concatenation, early fusion by concatenation, and early partial fusion. We conclude that 1) utilizing the AlexNet pre-trained parameters, 2) multi-scale image pyramid inputs, and 3) sharing network parameters contribute to better classification performance. We hope the work can inspire more studies for multi-scale analysis of WSI pathological images.

CHAPTER 6
CONCLUSION AND DISCUSSION

This dissertation has developed methods to enable precision medicine by integrating multi-modal biomedical data and integrating unlabeled or weakly labeled biomedical data. The proposed methods have been validated on multiple biomedical applications, including predicting Alzheimer's disease, multi-omics based breast cancer survival analysis, diagnosing Barrett's esophagus endomicroscopic images, and predicting breast cancer pathological images.

For independent multi-modal data, we utilize the complementary principle to integrate independent modalities by concatenating hidden features learned with independent feature representation. We have applied to the prediction of Alzheimer's Disease by integrating EHRs, SNPs, and MR images [116].

For dependent multi-modal data (e.g., multi-omics), we utilize the consensus principle to integrate dependent modalities by modeling the complex interactions implicitly and learning a modality-invariant feature representation. For the application of the proposed methods to multi-omics data, we first investigated RNA-seq pipeline selection for gene expression estimation [119]. We then applied cross-modality translation network [176] and divergence-based consensus network [183] to breast cancer multi-omics data.

We have developed an autoencoder-based semi-supervised learning framework for the generalized biomedical data integration for unlabeled data. We have applied the proposed network to improve Barrett's esophagus classification by integrating unlabeled endomicroscopic images [196]. We have also developed a multi-scale convolutional network to integrate the context-aware features for the image pyramids of whole-slide images and applied to breast cancer classification for improved performance [231].

In conclusion, we have developed multiple models to integrate multi-modal biomedical

data. By validating the proposed models to multiple biomedical applications, we have demonstrated the effectiveness and potential of multi-modal integration for biomedical data, which is essential towards precision medicine.

However, several limitations have been identified and can be improved to achieve more robust performance. The first limitation of the dissertation work is the feature extraction and selection for multi-omics data. With the "curse of dimensionality" in multi-omics data (i.e., small sample size vs. huge feature dimensions), feature selection or dimension reduction is essential for the success of machine learning techniques. In the dissertation, we have applied PCA or high-variance filtering to the multi-omics data for dimension reduction, which are two simple and unsupervised dimension reduction techniques. As a separate preprocessing step, the feature selection is performed on each data modality independently and separated from the downstream training. One future direction is to improve the feature extraction/selection for multi-omics data by end-to-end feature selection and predictive modeling.

The second limitation is the proposed multi-modality data integration approach. In this dissertation, we utilize consensus or complementary principles to integrated different multi-modal biomedical data. A promising direction would be to improve the data integration methods and explore strategies beyond consensus and complementary principle. One promising future direction could be causal inferences, in which we can identify causal relationships among modalities using multi-omics data. The causal inference among modalities can potentially facilitate knowledge mining and biological validation.

The third limitation of the methods developed in this dissertation is the lack of model interpretation. As a purely data-driven approach, the deep neural networks worked as a black box, and the model behaviors are hard to interpret or diagnose. However, model transparency is essential to build trust for caregivers and patients. The interpretation can also facilitate model validation, which can improve the model reproducibility and generalizability. Thus, one essential future direction is to enable model interpretation for the deep

learning-based multi-omics integration models and improve the model transparency.

The last limitation is the quality control of multi-modal biomedical data. Although basic quality controls have been applied to the multi-modal data utilized in this dissertation, a more robust and sophisticated quality control can significantly improve the stability of the proposed pipelines. The primary direction is to improve the modality-specific quality control (e.g., missing values for EHR data; color variations, motion artifacts, blur for medical images; sample contaminations and sequencing variations for multi-omics data). Another quality control direction is to mitigate missing modality issues by methods such as imputation, which can significantly increase the amount of training data.

REFERENCES

[1] P. C. Tang, J. S. Ash, D. W. Bates, J. M. Overhage, and D. Z. Sands, "Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption," *Journal of the American Medical Informatics Association*, vol. 13, no. 2, pp. 121–126, Mar. 1, 2006.

[2] B. G. a. S. Buchanan, "Rule- Based Expert Systems : The MYCIN Experiments of the Stanford Heuristic Programming Project," in *Xix, 748 p. Reading, Mass.: Addison-Wesley Pub. Co., 1984. Includes Bibliography: P. 717-738 and Subject Index*, CUMINCAD, 1984.

[3] R. S. Ledley and L. B. Lusted, "Reasoning Foundations of Medical Diagnosis," *Science*, vol. 130, no. 3366, pp. 9–21, 1959. JSTOR: 1758070.

[4] A. K. Jha, C. M. DesRoches, E. G. Campbell, K. Donelan, S. R. Rao, T. G. Ferris, A. Shields, S. Rosenbaum, and D. Blumenthal, "Use of Electronic Health Records in U.S. Hospitals," *New England Journal of Medicine*, vol. 360, no. 16, pp. 1628–1638, Apr. 16, 2009. pmid: 19321858.

[5] D. Charles, M. Gabriel, and M. F. Furukawa, "Adoption of Electronic Health Record Systems among U.S. Non-federal Acute Care Hospitals, 2008-2013," p. 9,

[6] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. G. Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher,

L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, "The Sequence of the Human Genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, Feb. 16, 2001. pmid: `11181995`.

[7] A. Ahmadian, B. Gharizadeh, A. C. Gustafsson, F. Sterky, P. Nyrén, M. Uhlén, and J. Lundeberg, "Single-Nucleotide Polymorphism Analysis by Pyrosequencing," *Analytical Biochemistry*, vol. 280, no. 1, pp. 103–110, Apr. 10, 2000.

[8] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 85–97, 2 Feb. 2006.

[9] A. J. Matlin, F. Clark, and C. W. J. Smith, "Understanding alternative splicing: Towards a cellular code," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 5, pp. 386–398, 5 May 2005.

[10] D. J. Hunter, "Gene–environment interactions in human diseases," *Nature Reviews Genetics*, vol. 6, no. 4, pp. 287–298, 4 Apr. 2005.

[11] B. Gülbakan, R. K. Özgül, A. Yüzbaşıoğlu, M. Kohl, H.-P. Deigner, and M. Özgüç, "Discovery of biomarkers in rare diseases: Innovative approaches by predictive and personalized medicine," *EPMA Journal*, vol. 7, no. 1, p. 24, Dec. 8, 2016.

[12] E. Check Hayden, "Technology: The $1,000 genome," *Nature News*, vol. 507, no. 7492, p. 294, Mar. 20, 2014.

[13] O. Bodenreider, "The Unified Medical Language System (UMLS): Integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, pp. D267–D270, suppl_1 Jan. 1, 2004.

[14] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 6 Jun. 2012.

[15] B. Middleton, M. Bloomrosen, M. A. Dente, B. Hashmat, R. Koppel, J. M. Overhage, T. H. Payne, S. T. Rosenbloom, C. Weaver, and J. Zhang, "Enhancing patient safety and quality of care by improving the usability of electronic health record systems: Recommendations from AMIA," *Journal of the American Medical Informatics Association*, vol. 20, no. e1, e2–e8, Jun. 1, 2013.

[16] J. Sun, F. Wang, J. Hu, and S. Edabollahi, "Supervised patient similarity measure of heterogeneous patient records," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 1, pp. 16–24, Dec. 10, 2012.

[17] C. G. Victora, J.-P. Habicht, and J. Bryce, "Evidence-Based Public Health: Moving Beyond Randomized Trials," *American Journal of Public Health*, vol. 94, no. 3, pp. 400–405, Mar. 1, 2004.

[18] E. A. Ashley, "The Precision Medicine Initiative: A New National Effort," *JAMA*, vol. 313, no. 21, p. 2119, Jun. 2, 2015.

[19] M.-C. King, J. H. Marks, and J. B. Mandell, "Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2," *Science*, vol. 302, no. 5645, pp. 643–646, Oct. 24, 2003. pmid: 14576434.

[20] E. D. Esplin, L. Oei, and M. P. Snyder, "Personalized sequencing and the future of medicine: Discovery, diagnosis and defeat of disease," *Pharmacogenomics*, vol. 15, no. 14, pp. 1771–1790, Nov. 2014. pmid: 25493570.

[21] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control," *Nature Medicine*, vol. 10, no. 8, pp. 789–799, 8 Aug. 2004.

[22] R. L. Strausberg, A. J. G. Simpson, L. J. Old, and G. J. Riggins, "Oncogenomics and the development of new cancer therapies," *Nature*, vol. 429, no. 6990, pp. 469–474, 6990 May 2004.

[23] P. Nguyen, A. Taghian, M. Katz, A. Niemierko, R. Raad, W. Boon, J. Bellon, J. Wong, B. Smith, and J. Harris, "Breast Cancer Subtype Approximated by Estrogen Receptor, Progesterone Receptor, and HER-2 Is Associated With Local and Distant Recurrence After Breast-Conserving Therapy," *Journal of clinical oncology :*

official journal of the American Society of Clinical Oncology, vol. 26, pp. 2373–8, Jun. 1, 2008.

[24] S. L. Zheng, J. Sun, F. Wiklund, S. Smith, P. Stattin, G. Li, H.-O. Adami, F.-C. Hsu, Y. Zhu, K. Bälter, A. K. Kader, A. R. Turner, W. Liu, E. R. Bleecker, D. A. Meyers, D. Duggan, J. D. Carpten, B.-L. Chang, W. B. Isaacs, J. Xu, and H. Grönberg, "Cumulative Association of Five Genetic Variants with Prostate Cancer," *New England Journal of Medicine*, vol. 358, no. 9, pp. 910–919, Feb. 28, 2008. pmid: 18199855.

[25] P. B. Chapman, A. Hauschild, C. Robert, J. B. Haanen, P. Ascierto, J. Larkin, R. Dummer, C. Garbe, A. Testori, M. Maio, D. Hogg, P. Lorigan, C. Lebbe, T. Jouary, D. Schadendorf, A. Ribas, S. J. O'Day, J. A. Sosman, J. M. Kirkwood, A. M. Eggermont, B. Dreno, K. Nolop, J. Li, B. Nelson, J. Hou, R. J. Lee, K. T. Flaherty, and G. A. McArthur, "Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation," *New England Journal of Medicine*, vol. 364, no. 26, pp. 2507–2516, Jun. 30, 2011. pmid: 21639808.

[26] R. M. Durbin, D. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, F. S. Collins, F. M. De La Vega, P. Donnelly, M. Egholm, P. Flicek, S. B. Gabriel, R. A. Gibbs, B. M. Knoppers, E. S. Lander, H. Lehrach, E. R. Mardis, G. A. McVean, D. A. Nickerson, L. Peltonen, A. J. Schafer, S. T. Sherry, J. Wang, R. K. Wilson, R. A. Gibbs, D. Deiros, M. Metzker, D. Muzny, J. Reid, D. Wheeler, J. Wang, J. Li, M. Jian, G. Li, R. Li, H. Liang, G. Tian, B. Wang, J. Wang, W. Wang, H. Yang, X. Zhang, H. Zheng, E. S. Lander, D. Altshuler, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, S. B. Gabriel, D. B. Jaffe, E. Shefler, C. L. Sougnez, D. R. Bentley, N. Gormley, S. Humphray, Z. Kingsbury, P. Kokko-Gonzales, J. Stone, K. J. McKernan, G. L. Costa, J. K. Ichikawa, C. C. Lee, R. Sudbrak, H. Lehrach, T. A. Borodina, A. Dahl, A. N. Davydov, P. Marquardt, F. Mertes, W. Nietfeld, P. Rosenstiel, S. Schreiber, A. V. Soldatov, B. Timmermann, M. Tolzmann, M. Egholm, J. Affourtit, D. Ashworth, S. Attiya, M. Bachorski, E. Buglione, A. Burke, A. Caprio, C. Celone, S. Clark, D. Conners, B. Desany, L. Gu, L. Guccione, K. Kao, A. Kebbel, J. Knowlton, M. Labrecque, L. McDade, C. Mealmaker, M. Minderman, A. Nawrocki, F. Niazi, K. Pareja, R. Ramenani, D. Riches, W. Song, C. Turcotte, S. Wang, E. R. Mardis, R. K. Wilson, D. Dooling, L. Fulton, R. Fulton, G. Weinstock, R. M. Durbin, J. Burton, D. M. Carter, C. Churcher, A. Coffey, A. Cox, A. Palotie, M. Quail, T. Skelly, J. Stalker, H. P. Swerdlow, D. Turner, A. De Witte, S. Giles, R. A. Gibbs, D. Wheeler, M. Bainbridge, D. Challis, A. Sabo, F. Yu, J. Yu, J. Wang, X. Fang, X. Guo, R. Li, Y. Li, R. Luo, S. Tai, H. Wu, H. Zheng, X. Zheng, Y. Zhou, G. Li, J. Wang, H. Yang, G. T. Marth, E. P. Garrison, W. Huang, A. Indap, D. Kural, W.-P. Lee, W. Fung Leong, A. R. Quinlan, C. Stewart, M. P. Stromberg, A. N. Ward, J. Wu, C. Lee, R. E. Mills, X. Shi, M. J. Daly, M. A. DePristo, D. Altshuler, A. D. Ball, E. Banks, T. Bloom, B. L. Browning, K. Cibulskis, T. J. Fennell, K. V. Garimella, S. R. Grossman, R. E. Handsaker, M. Hanna, C. Hartl, D. B. Jaffe, A. M. Kernytsky, J. M. Korn, H. Li, J. R. Maguire,

S. A. McCarroll, A. McKenna, J. C. Nemesh, A. A. Philippakis, R. E. Poplin, A. Price, M. A. Rivas, P. C. Sabeti, S. F. Schaffner, E. Shefler, I. A. Shlyakhter, D. N. Cooper, E. V. Ball, M. Mort, A. D. Phillips, P. D. Stenson, J. Sebat, V. Makarov, K. Ye, S. C. Yoon, C. D. Bustamante, L. Clarke, P. Flicek, F. Cunningham, J. Herrero, S. Keenen, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, R. E. Smith, V. Zalunin, X. Zheng-Bradley, J. O. Korbel, A. M. Stütz, S. Humphray, M. Bauer, R. Keira Cheetham, T. Cox, M. Eberle, T. James, S. Kahn, L. Murray, A. Chakravarti, K. Ye, F. M. De La Vega, Y. Fu, F. C. L. Hyland, J. M. Manning, S. F. McLaughlin, H. E. Peckham, O. Sakarya, Y. A. Sun, E. F. Tsung, M. A. Batzer, M. K. Konkel, J. A. Walker, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, R. Herwig, D. V. Parkhomchuk, S. T. Sherry, R. Agarwala, H. M. Khouri, A. O. Morgulis, J. E. Paschall, L. D. Phan, K. E. Rotmistrovsky, R. D. Sanders, M. F. Shumway, C. Xiao, G. A. McVean, A. Auton, Z. Iqbal, G. Lunter, J. L. Marchini, L. Moutsianas, S. Myers, A. Tumian, B. Desany, J. Knight, R. Winer, D. W. Craig, S. M. Beckstrom-Sternberg, A. Christoforides, The 1000 Genomes Project Consortium, Corresponding author, Steering committee, Production group: Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Illumina, Life Technologies, Max Planck Institute for Molecular Genetics, Roche Applied Science, Washington University in St Louis, Wellcome Trust Sanger Institute, Analysis group: Agilent Technologies, Baylor College of Medicine, Boston College, Brigham and Women's Hospital, T. H. G. M. D. Cardiff University, Cold Spring Harbor Laboratory, Cornell and Stanford Universities, European Bioinformatics Institute, European Molecular Biology Laboratory, Johns Hopkins University, Leiden University Medical Center, Louisiana State University, US National Institutes of Health, Oxford University, and The Translational Genomics Research Institute, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 7319 Oct. 2010.

[27] A. Telenti, L. C. T. Pierce, W. H. Biggs, J. di Iulio, E. H. M. Wong, M. M. Fabani, E. F. Kirkness, A. Moustafa, N. Shah, C. Xie, S. C. Brewerton, N. Bulsara, C. Garner, G. Metzker, E. Sandoval, B. A. Perkins, F. J. Och, Y. Turpaz, and J. C. Venter, "Deep sequencing of 10,000 human genomes," *Proceedings of the National Academy of Sciences*, vol. 113, no. 42, pp. 11 901–11 906, Oct. 18, 2016.

[28] J. R. Ecker, W. A. Bickmore, I. Barroso, J. K. Pritchard, Y. Gilad, and E. Segal, "ENCODE explained," *Nature*, vol. 489, no. 7414, pp. 52–54, 7414 Sep. 2012.

[29] N. Allen, C. Sudlow, P. Downey, T. Peakman, J. Danesh, P. Elliott, J. Gallacher, J. Green, P. Matthews, J. Pell, T. Sprosen, and R. Collins, "UK Biobank: Current status and what it means for epidemiology," *Health Policy and Technology*, vol. 1, no. 3, pp. 123–126, Sep. 1, 2012.

[30] M. Gerlinger, A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N. Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. R. Santos, M.

Nohadani, A. C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P. A. Futreal, and C. Swanton, "Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing," *New England Journal of Medicine*, vol. 366, no. 10, pp. 883–892, Mar. 8, 2012. pmid: 22397650.

[31] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The Alzheimer's Disease Neuroimaging Initiative," *Neuroimaging Clinics*, vol. 15, no. 4, pp. 869–877, Nov. 1, 2005.

[32] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, W. Poewe, B. Mollenhauer, P.-E. Klinik, T. Sherer, M. Frasier, C. Meunier, A. Rudolph, C. Casaceli, J. Seibyl, S. Mendick, N. Schuff, Y. Zhang, A. Toga, K. Crawford, A. Ansbach, P. De Blasio, M. Piovella, J. Trojanowski, L. Shaw, A. Singleton, K. Hawkins, J. Eberling, D. Brooks, D. Russell, L. Leary, S. Factor, B. Sommerfeld, P. Hogarth, E. Pighetti, K. Williams, D. Standaert, S. Guthrie, R. Hauser, H. Delgado, J. Jankovic, C. Hunter, M. Stern, B. Tran, J. Leverenz, M. Baca, S. Frank, C.-A. Thomas, I. Richard, C. Deeley, L. Rees, F. Sprenger, E. Lang, H. Shill, S. Obradov, H. Fernandez, A. Winters, D. Berg, K. Gauss, D. Galasko, D. Fontaine, Z. Mari, M. Gerstenhaber, D. Brooks, S. Malloy, P. Barone, K. Longo, T. Comery, B. Ravina, I. Grachev, K. Gallagher, M. Collins, K. L. Widnell, S. Ostrowizki, P. Fontoura, T. Ho, J. Luthman, M. van der Brug, A. D. Reith, and P. Taylor, "The Parkinson Progression Marker Initiative (PPMI)," *Progress in Neurobiology*, Biological Markers for Neurodegenerative Diseases, vol. 95, no. 4, pp. 629–635, Dec. 1, 2011.

[33] S. R. Horbal, W. Seffens, A. R. Davis, N. Silvestrov, G. H. Gibbons, R. C. Quarells, and A. Bidulescu, "Associations of Apelin, Visfatin, and Urinary 8-Isoprostane With Severe Hypertension in African Americans: The MH-GRID Study," *American Journal of Hypertension*, vol. 29, no. 7, pp. 814–820, Jul. 1, 2016.

[34] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nature Methods*, vol. 5, no. 1, pp. 16–18, 1 Jan. 2008.

[35] S. Kothari, J. H. Phan, T. H. Stokes, and M. D. Wang, "Pathology imaging informatics for quantitative analysis of whole-slide images," *Journal of the American Medical Informatics Association*, vol. 20, no. 6, pp. 1099–1108, Nov. 1, 2013.

[36] S. Kothari, J. H. Phan, A. O. Osunkoya, and M. D. Wang, "Biological interpretation of morphological patterns in histopathological whole-slide images," in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, ser. BCB '12, New York, NY, USA: Association for Computing Machinery, Oct. 7, 2012, pp. 218–225, ISBN: 978-1-4503-1670-5.

[37] J. H. Phan, C. F. Quo, C. Cheng, and M. D. Wang, "Multiscale Integration of -Omic, Imaging, and Clinical Data in Biomedical Informatics," *IEEE Reviews in Biomedical Engineering*, vol. 5, pp. 74–87, 2012.

[38] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1240–1251, May 2016.

[39] D. Kumar, A. Wong, and D. A. Clausi, "Lung Nodule Classification Using Deep Features in CT Images," in *2015 12th Conference on Computer and Robot Vision*, Jun. 2015, pp. 133–138.

[40] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, "Classification of breast cancer histology images using Convolutional Neural Networks," *PLOS ONE*, vol. 12, no. 6, e0177544, Jun. 1, 2017.

[41] K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin, and M. Snyder, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature Communications*, vol. 7, no. 1, p. 12 474, 1 Aug. 16, 2016.

[42] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2424–2433.

[43] L. Tong, R. Hoffman, S. R. Deshpande, and M. D. Wang, "Predicting heart rejection using histopathological whole-slide imaging and deep neural network with dropout," in *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, Feb. 2017, pp. 1–4.

[44] L. Wong, "Technologies for integrating biological data," *Briefings in Bioinformatics*, vol. 3, no. 4, pp. 389–404, Dec. 1, 2002.

[45] C. Goble and R. Stevens, "State of the nation in data integration for bioinformatics," *Journal of Biomedical Informatics*, Semantic Mashup of Biomedical Data, vol. 41, no. 5, pp. 687–693, Oct. 1, 2008.

[46] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegnér, "Data integration in the era of omics: Current and future challenges," *BMC Systems Biology*, vol. 8, no. 2, p. I1, Mar. 13, 2014.

[47] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype–phenotype interactions," *Nature Reviews Genetics*, vol. 16, no. 2, pp. 85–97, 2 Feb. 2015.

[48] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, "Real-Time DNA Sequencing from Single Polymerase Molecules," *Science*, vol. 323, no. 5910, pp. 133–138, Jan. 2, 2009. pmid: 19023044.

[49] D. J. Brailer, "Interoperability: The Key To The Future Health Care System," *Health Affairs*, vol. 24, W5–19, Suppl1 Jan. 1, 2005.

[50] R. H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F. M. Behlen, P. V. Biron, and A. Shabo (Shvo), "HL7 Clinical Document Architecture, Release 2," *Journal of the American Medical Informatics Association*, vol. 13, no. 1, pp. 30–39, Jan. 1, 2006.

[51] G. Alterovitz, J. Warner, P. Zhang, Y. Chen, M. Ullman-Cullere, D. Kreda, and I. S. Kohane, "SMART on FHIR Genomics: Facilitating standardized clinico-genomic apps," *Journal of the American Medical Informatics Association*, vol. 22, no. 6, pp. 1173–1178, Nov. 1, 2015.

[52] L. Shi, L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Scherf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T.-M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X.-h. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q.-Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J.

Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, W. Slikker, and MAQC Consortium, "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology*, vol. 24, no. 9, pp. 1151–1161, 9 Sep. 2006.

[53] SEQC/MAQC-III Consortium, "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium," *Nature Biotechnology*, vol. 32, no. 9, pp. 903–914, Sep. 2014.

[54] L. Chin, J. N. Andersen, and P. A. Futreal, "Cancer genomics: From discovery science to personalized medicine," *Nature Medicine*, vol. 17, no. 3, pp. 297–303, Mar. 2011.

[55] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, Jul. 1959.

[56] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 7553 May 2015.

[57] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[58] C. Clifton, M. Kantarcioğlu, A. Doan, G. Schadow, J. Vaidya, A. Elmagarmid, and D. Suciu, "Privacy-preserving data integration and sharing," in *Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery - DMKD '04*, Paris, France: ACM Press, 2004, p. 19, ISBN: 978-1-58113-908-2.

[59] D. Bender and K. Sartipi, "HL7 FHIR: An Agile and RESTful approach to healthcare information exchange," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, Jun. 2013, pp. 326–331.

[60] B. Franz, "Applying FHIR in an Integrated Health Monitoring System," *European Journal for Biomedical Informatics*, vol. 11, no. 02, 2015.

[61] L. Meier, S. V. D. Geer, and P. Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[62] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 03, no. 02, pp. 185–205, Apr. 1, 2005.

[63] M. Gonen, E. Alpaydın, B. E. Tr, and B. E. Tr, "Multiple Kernel Learning Algorithms," p. 58, 2011.

[64] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, Dec. 1, 2004.

[65] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009, 1268 pp., ISBN: 978-0-262-01319-2. Google Books: 7dzpHCHzNQ4C.

[66] N. Srivastava and R. Salakhutdinov, "Multimodal Learning with Deep Boltzmann Machines," p. 32,

[67] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep Canonical Correlation Analysis," in *International Conference on Machine Learning*, PMLR, May 26, 2013, pp. 1247–1255.

[68] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "HeMIS: Hetero-Modal Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 469–477, ISBN: 978-3-319-46723-8.

[69] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, New York, NY, USA: Association for Computing Machinery, Aug. 13, 2016, pp. 1135–1144, ISBN: 978-1-4503-4232-2.

[70] P. W. Koh and P. Liang. (Jul. 9, 2017). Understanding Black-box Predictions via Influence Functions. arXiv: 1703.04730 [cs, stat], (visited on 11/03/2020).

[71] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems*, ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2000, pp. 1–15, ISBN: 978-3-540-45014-6.

[72] H.-B. Shen and K.-C. Chou, "Ensemble classifier for protein fold pattern recognition," *Bioinformatics*, vol. 22, no. 14, pp. 1717–1722, Jul. 15, 2006.

[73] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1, 1996.

[74] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. JSTOR: 2699986.

[75] U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. A. Garraway, and D. Pe'er, "An Integrated Approach to Uncover Drivers of Cancer," *Cell*, vol. 143, no. 6, pp. 1005–1017, Dec. 10, 2010.

[76] O. Gevaert, F. D. Smet, D. Timmerman, Y. Moreau, and B. D. Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks," *Bioinformatics*, vol. 22, no. 14, e184–e190, Jul. 15, 2006.

[77] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10, New York, NY, USA: Association for Computing Machinery, Oct. 25, 2010, pp. 251–260, ISBN: 978-1-60558-933-6.

[78] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, "Counterfactual reasoning and learning systems: The example of computational advertising," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3207–3260, Jan. 1, 2013.

[79] Y. EL-Manzalawy, T.-Y. Hsieh, M. Shivakumar, D. Kim, and V. Honavar, "Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data," *BMC Medical Genomics*, vol. 11, no. 3, p. 71, Sep. 14, 2018.

[80] O. B. Poirion, K. Chaudhary, and L. X. Garmire, "Deep Learning data integration for better risk stratification models of bladder cancer," *AMIA Summits on Translational Science Proceedings*, vol. 2018, pp. 197–206, May 18, 2018. pmid: 29888072.

[81] T. Ma and A. Zhang, "Multi-view Factorization AutoEncoder with Network Constraints for Multi-omic Integrative Analysis," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2018, pp. 702–707.

[82] Z. Huang, X. Zhan, S. Xiang, T. S. Johnson, B. Helm, C. Y. Yu, J. Zhang, P. Salama, M. Rizkalla, Z. Han, and K. Huang, "SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer," *Frontiers in Genetics*, vol. 10, 2019.

[83] J. Mitchel, K. Chatlin, L. Tong, and M. D. Wang, "A Translational Pipeline for Overall Survival Prediction of Breast Cancer Patients by Decision-Level Integration of Multi-Omics Data," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2019, pp. 1573–1580.

[84] C. Xu, D. Tao, and C. Xu. (Apr. 20, 2013). A Survey on Multi-view Learning. arXiv: 1304.5634 [cs], (visited on 01/10/2020).

[85] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. (Aug. 1, 2017). Multimodal Machine Learning: A Survey and Taxonomy. arXiv: 1705.09406 [cs], (visited on 01/10/2020).

[86] Y. Song, L.-P. Morency, and R. Davis, "Multi-view latent variable discriminative models for action recognition," presented at the 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2120–2127, ISBN: 1-4673-1228-2.

[87] K. Liu, Y. Li, N. Xu, and P. Natarajan. (May 29, 2018). Learn to Combine Modalities in Multimodal Deep Learning. arXiv: 1805.11730 [cs, stat], (visited on 11/18/2019).

[88] W. Guo, J. Wang, and S. Wang, "Deep Multimodal Representation Learning: A Survey," *IEEE Access*, vol. 7, pp. 63 373–63 394, 2019.

[89] Y. Li, F.-X. Wu, and A. Ngom, "A review on machine learning principles for multi-view biological data integration," *Briefings in Bioinformatics*, vol. 19, no. 2, pp. 325–340, Mar. 1, 2018.

[90] M. Kim and I. Tagkopoulos, "Data integration and predictive modeling methods for multi-omics datasets," *Molecular Omics*, vol. 14, no. 1, pp. 8–25, Feb. 12, 2018.

[91] A. Xu, J. Chen, H. Peng, G. Han, and H. Cai, "Simultaneous Interrogation of Cancer Omics to Identify Subtypes With Significant Clinical Differences," *Frontiers in Genetics*, vol. 10, 2019.

[92] S. V. Vasaikar, P. Straub, J. Wang, and B. Zhang, "LinkedOmics: Analyzing multi-omics data within and across 32 cancer types," *Nucleic Acids Research*, vol. 46, no. D1, pp. D956–D963, Jan. 4, 2018.

[93] G. P. Way, F. Sanchez-Vega, K. La, J. Armenia, W. K. Chatila, A. Luna, C. Sander, A. D. Cherniack, M. Mina, and G. Ciriello, "Machine learning detects pan-cancer ras pathway activation in the cancer genome atlas," *Cell reports*, vol. 23, no. 1, pp. 172–180, 2018.

[94] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, "Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer," *Clinical Cancer Research*, vol. 24, no. 6, pp. 1248–1259, Mar. 15, 2018. pmid: 28982688.

[95] R. J. Perrin, A. M. Fagan, and D. M. Holtzman, "Multimodal techniques for diagnosis and prognosis of Alzheimer's disease," *Nature*, vol. 461, no. 7266, pp. 916–922, 7266 Oct. 2009.

[96] K. Blennow, B. Dubois, A. M. Fagan, P. Lewczuk, M. J. de Leon, and H. Hampel, "Clinical utility of cerebrospinal fluid biomarkers in the diagnosis of early Alzheimer's disease," *Alzheimer's & Dementia*, vol. 11, no. 1, pp. 58–69, Jan. 1, 2015.

[97] S. F. Eskildsen, P. Coupé, V. S. Fonov, J. C. Pruessner, and D. L. Collins, "Structural imaging biomarkers of Alzheimer's disease: Predicting disease progression," *Neurobiology of Aging*, Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders (NIBAD), vol. 36, S23–S31, Jan. 1, 2015.

[98] T. Grimmer, C. Wutz, P. Alexopoulos, A. Drzezga, S. Förster, H. Förstl, O. Goldhardt, M. Ortner, C. Sorg, and A. Kurz, "Visual Versus Fully Automated Analyses of 18F-FDG and Amyloid PET for Prediction of Dementia Due to Alzheimer Disease in Mild Cognitive Impairment," *Journal of Nuclear Medicine*, vol. 57, no. 2, pp. 204–207, Feb. 1, 2016. pmid: 26585056.

[99] R. Cui and M. Liu, "RNN-based longitudinal analysis for diagnosis of Alzheimer's disease," *Computerized Medical Imaging and Graphics*, vol. 73, pp. 1–10, Apr. 1, 2019.

[100] J. Barnes, O. T. Carmichael, K. K. Leung, C. Schwarz, G. R. Ridgway, J. W. Bartlett, I. B. Malone, J. M. Schott, M. N. Rossor, G. J. Biessels, C. DeCarli, and N. C. Fox, "Vascular and Alzheimer's disease markers independently predict brain atrophy rate in Alzheimer's Disease Neuroimaging Initiative controls," *Neurobiology of Aging*, vol. 34, no. 8, pp. 1996–2002, Aug. 1, 2013.

[101] J. D. Doecke, "Blood-Based Protein Biomarkers for Diagnosis of Alzheimer Disease," *Archives of Neurology*, vol. 69, no. 10, p. 1318, Oct. 1, 2012.

[102] G. Lee, K. Nho, B. Kang, K.-A. Sohn, and D. Kim, "Predicting Alzheimer's disease progression using multi-modal deep learning approach," *Scientific Reports*, vol. 9, no. 1, p. 1952, 1 Feb. 13, 2019.

[103] J. Zhao, Q. Feng, P. Wu, R. A. Lupu, R. A. Wilke, Q. S. Wells, J. C. Denny, and W.-Q. Wei, "Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction," *Scientific Reports*, vol. 9, no. 1, p. 717, 1 Jan. 24, 2019.

[104] M. Dyrba, M. Grothe, T. Kirste, and S. J. Teipel, "Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM," *Human Brain Mapping*, vol. 36, no. 6, pp. 2118–2131, 2015.

[105] J. L. Shaffer, J. R. Petrella, F. C. Sheldon, K. R. Choudhury, V. D. Calhoun, R. E. Coleman, and P. M. Doraiswamy, "Predicting Cognitive Decline in Subjects at Risk

for Alzheimer Disease by Using Combined Cerebrospinal Fluid, MR Imaging, and PET Biomarkers," *Radiology*, vol. 266, no. 2, pp. 583–591, Feb. 1, 2013.

[106] Z. Dai, C. Yan, Z. Wang, J. Wang, M. Xia, K. Li, and Y. He, "Discriminative analysis of early Alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (M3)," *NeuroImage*, vol. 59, no. 3, pp. 2187–2195, Feb. 1, 2012.

[107] M. Dyrba, F. Barkhof, A. Fellgiebel, M. Filippi, L. Hausner, K. Hauenstein, T. Kirste, and S. J. Teipel, "Predicting Prodromal Alzheimer's Disease in Subjects with Mild Cognitive Impairment Using Machine Learning Classification of Multimodal Multicenter Diffusion-Tensor and Magnetic Resonance Imaging Data," *Journal of Neuroimaging*, vol. 25, no. 5, pp. 738–747, 2015.

[108] M. Lorenzi, I. J. Simpson, A. F. Mendelson, S. B. Vos, M. J. Cardoso, M. Modat, J. M. Schott, and S. Ourselin, "Multimodal Image Analysis in Alzheimer's Disease via Statistical Modelling of Non-local Intensity Correlations," *Scientific Reports*, vol. 6, no. 1, p. 22 161, 1 Apr. 11, 2016.

[109] J. W. Vogel, E. Vachon-Presseau, A. Pichet Binette, A. Tam, P. Orban, R. La Joie, M. Savard, C. Picard, J. Poirier, P. Bellec, J. C. S. Breitner, and S. Villeneuve, "Brain properties predict proximity to symptom onset in sporadic Alzheimer's disease," *Brain*, vol. 141, no. 6, pp. 1871–1883, Jun. 1, 2018.

[110] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, and D. Rueckert, "Random forest-based similarity measures for multi-modal classification of Alzheimer's disease," *NeuroImage*, vol. 65, pp. 167–175, Jan. 15, 2013.

[111] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, Apr. 1, 2011.

[112] H. Wang, F. Nie, H. Huang, S. L. Risacher, A. J. Saykin, L. Shen, and F. t. A. D. N. Initiative, "Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning," *Bioinformatics*, vol. 28, no. 12, pp. i127–i136, Jun. 15, 2012.

[113] H.-I. Suk, S.-W. Lee, D. Shen, and The Alzheimer's Disease Neuroimaging Initiative, "Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis," *Brain Structure and Function*, vol. 221, no. 5, pp. 2569–2587, Jun. 1, 2016.

[114] H.-I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, Nov. 1, 2014.

[115] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI)," *Alzheimer's & Dementia*, vol. 1, no. 1, pp. 55–66, Jul. 1, 2005.

[116] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of alzheimer's disease stage," *Scientific Reports*, 2020.

[117] L. Shen, P. M. Thompson, S. G. Potkin, L. Bertram, L. A. Farrer, T. M. Foroud, R. C. Green, X. Hu, M. J. Huentelman, S. Kim, J. S. K. Kauwe, Q. Li, E. Liu, F. Macciardi, J. H. Moore, L. Munsie, K. Nho, V. K. Ramanan, S. L. Risacher, D. J. Stone, S. Swaminathan, A. W. Toga, M. W. Weiner, A. J. Saykin, and for the Alzheimer's Disease Neuroimaging Initiative, "Genetic analysis of quantitative phenotypes in AD and MCI: Imaging, cognition and biomarkers," *Brain Imaging and Behavior*, vol. 8, no. 2, pp. 183–207, Jun. 1, 2014.

[118] S. Leandrou, S. Petroudi, P. A. Kyriacou, C. C. Reyes-Aldasoro, and C. S. Pattichis, "Quantitative MRI Brain Studies in Mild Cognitive Impairment and Alzheimer's Disease: A Methodological Review," *IEEE Reviews in Biomedical Engineering*, vol. 11, pp. 97–111, 2018.

[119] L. Tong, P.-Y. Wu, J. H. Phan, H. R. Hassazadeh, W. Tong, and M. D. Wang, "Impact of RNA-seq data analysis algorithms on gene expression estimation and downstream prediction," *Scientific Reports*, vol. 10, no. 1, p. 17 925, 1 Oct. 21, 2020.

[120] "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models," *Nature Biotechnology*, vol. 28, no. 8, pp. 827–838, Aug. 2010. pmid: 20676074.

[121] F. Ozsolak and P. M. Milos, "RNA sequencing: Advances, challenges and opportunities," *Nature Reviews Genetics*, vol. 12, no. 2, pp. 87–98, 2 Feb. 2011.

[122] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: A revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 1 Jan. 2009.

[123] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, vol. 18, no. 9, pp. 1509–1517, Sep. 1, 2008. pmid: 18550803.

[124] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 7 Jul. 2008.

[125] J. Costa-Silva, D. Domingues, and F. M. Lopes, "RNA-Seq differential expression analysis: An extended review and a software tool," *PLOS ONE*, vol. 12, no. 12, e0190152, Dec. 21, 2017.

[126] C. R. Williams, A. Baccarella, J. Z. Parrish, and C. C. Kim, "Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq," *BMC Bioinformatics*, vol. 18, no. 1, p. 38, Jan. 17, 2017.

[127] G. Rigaill, S. Balzergue, V. Brunaud, E. Blondet, A. Rau, O. Rogier, J. Caius, C. Maugis-Rabusseau, L. Soubigou-Taconnat, S. Aubourg, C. Lurin, M.-L. Martin-Magniette, and E. Delannoy, "Synthetic data sets for the identification of key ingredients for RNA-seq differential analysis," *Briefings in Bioinformatics*, vol. 19, no. 1, pp. 65–76, Jan. 1, 2018.

[128] G. A. Merino, A. Conesa, and E. A. Fernández, "A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies," *Briefings in Bioinformatics*, vol. 20, no. 2, pp. 471–481, Mar. 25, 2019.

[129] M. Dapas, M. Kandpal, Y. Bi, and R. V. Davuluri, "Comparative evaluation of isoform-level gene expression estimation algorithms for RNA-seq and exon-array platforms," *Briefings in Bioinformatics*, vol. 18, no. 2, pp. 260–269, Mar. 1, 2017.

[130] A. Kanitz, F. Gypas, A. J. Gruber, A. R. Gruber, G. Martin, and M. Zavolan, "Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data," *Genome Biology*, vol. 16, no. 1, p. 150, Jul. 23, 2015.

[131] S. M. E. Sahraeian, M. Mohiyuddin, R. Sebra, H. Tilgner, P. T. Afshar, K. F. Au, N. Bani Asadi, M. B. Gerstein, W. H. Wong, M. P. Snyder, E. Schadt, and H. Y. K. Lam, "Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis," *Nature Communications*, vol. 8, no. 1, p. 59, 1 Jul. 5, 2017.

[132] S. Kumar, A. D. Vo, F. Qin, and H. Li, "Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data," *Scientific Reports*, vol. 6, no. 1, p. 21 597, 1 Feb. 10, 2016.

[133] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, "A survey of best practices for RNA-seq data analysis," *Genome Biology*, vol. 17, no. 1, p. 13, Jan. 26, 2016.

[134] L. Peixoto, D. Risso, S. G. Poplawski, M. E. Wimmer, T. P. Speed, M. A. Wood, and T. Abel, "How data analysis affects power, reproducibility and biological insight of

RNA-seq studies in complex datasets," *Nucleic Acids Research*, vol. 43, no. 16, pp. 7664–7674, Sep. 18, 2015.

[135] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, R25, Mar. 4, 2009.

[136] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, Apr. 2012.

[137] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, Jul. 15, 2009.

[138] T. D. Wu and S. Nacu, "Fast and SNP-tolerant detection of complex variants and splicing in short reads," *Bioinformatics*, vol. 26, no. 7, pp. 873–881, Apr. 1, 2010.

[139] D. Thierry-Mieg and J. Thierry-Mieg, "AceView: A comprehensive cDNA-supported gene and transcripts annotation," *Genome Biology*, vol. 7, no. 1, S12, Aug. 7, 2006.

[140] K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, J. N. MacLeod, D. Y. Chiang, J. F. Prins, and J. Liu, "MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery," *Nucleic Acids Research*, vol. 38, no. 18, e178–e178, Oct. 1, 2010.

[141] J. Hu, H. Ge, M. Newman, and K. Liu, "OSA: A fast and accurate alignment tool for RNA-Seq," *Bioinformatics*, vol. 28, no. 14, pp. 1933–1934, Jul. 15, 2012.

[142] G. R. Grant, M. H. Farkas, A. D. Pizarro, N. F. Lahens, J. Schug, B. P. Brunk, C. J. Stoeckert, J. B. Hogenesch, and E. A. Pierce, "Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)," *Bioinformatics*, vol. 27, no. 18, pp. 2518–2528, Sep. 15, 2011.

[143] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: Ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 1, 2013.

[144] Y. Liao, G. K. Smyth, and W. Shi, "The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote," *Nucleic Acids Research*, vol. 41, no. 10, e108–e108, May 1, 2013.

[145] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: Discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, May 1, 2009.

[146] Y. Li, J. M. Patel, and A. Terrell, "WHAM: A High-Throughput Sequence Alignment Method," *ACM Transactions on Database Systems*, vol. 37, no. 4, 28:1–28:39, Dec. 1, 2012.

[147] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," *Nature Methods*, vol. 8, no. 6, pp. 469–477, 6 Jun. 2011.

[148] S. Anders, P. T. Pyl, and W. Huber, "HTSeq—a Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, Jan. 15, 2015.

[149] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, 5 May 2010.

[150] B. Li and C. N. Dewey, "RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, no. 1, p. 323, Aug. 4, 2011.

[151] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, and F. Jaffrézic, "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis," *Briefings in Bioinformatics*, vol. 14, no. 6, pp. 671–683, Nov. 1, 2013.

[152] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Nature Precedings*, pp. 1–1, Apr. 30, 2010.

[153] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biology*, vol. 11, no. 3, R25, Mar. 2, 2010.

[154] S. A. Hardwick, I. W. Deveson, and T. R. Mercer, "Reference standards for next-generation sequencing," *Nature Reviews Genetics*, vol. 18, no. 8, pp. 473–484, Aug. 2017.

[155] R. Lindner and C. C. Friedel, "A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq," *PLOS ONE*, vol. 7, no. 12, e52403, Dec. 26, 2012.

[156] P. G. Engström, T. Steijger, B. Sipos, G. R. Grant, A. Kahles, G. Rätsch, N. Goldman, T. J. Hubbard, J. Harrow, R. Guigó, and P. Bertone, "Systematic evaluation of

spliced alignment programs for RNA-seq data," *Nature Methods*, vol. 10, no. 12, pp. 1185–1191, 12 Dec. 2013.

[157] A. Hatem, D. Bozdağ, A. E. Toland, and V. Çatalyürek, "Benchmarking short sequence mapping tools," *BMC Bioinformatics*, vol. 14, no. 1, p. 184, Jun. 7, 2013.

[158] I. Borozan, S. N. Watt, and V. Ferretti, "Evaluation of Alignment Algorithms for Discovery and Identification of Pathogens Using RNA-Seq," *PLOS ONE*, vol. 8, no. 10, e76935, Oct. 30, 2013.

[159] G. Baruzzo, K. E. Hayer, E. J. Kim, B. Di Camillo, G. A. FitzGerald, and G. R. Grant, "Simulation-based comprehensive benchmarking of RNA-seq aligners," *Nature Methods*, vol. 14, no. 2, pp. 135–139, Feb. 2017.

[160] C. Yang, P.-Y. Wu, L. Tong, J. Phan, and M. Wang, "The impact of RNA-seq aligners on gene expression estimation," in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, ser. BCB '15, New York, NY, USA: Association for Computing Machinery, Sep. 9, 2015, pp. 462–471, ISBN: 978-1-4503-3853-0.

[161] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nature Biotechnology*, vol. 34, no. 5, pp. 525–527, 5 May 2016.

[162] E. Maza, P. Frasse, P. Senin, M. Bouzayen, and M. Zouine, "Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments," *Communicative & Integrative Biology*, vol. 6, no. 6, e25849, Nov. 9, 2013. pmid: 26442135.

[163] H. Aanes, C. Winata, L. F. Moen, O. Østrup, S. Mathavan, P. Collas, T. Rognes, and P. Aleström, "Normalization of RNA-Sequencing Data from Samples with Varying mRNA Levels," *PLOS ONE*, vol. 9, no. 2, e89158, Feb. 25, 2014.

[164] N. A. Fonseca, J. Marioni, and A. Brazma, "RNA-Seq Gene Profiling - A Systematic Empirical Comparison," *PLOS ONE*, vol. 9, no. 9, e107026, Sep. 30, 2014.

[165] I. Nookaew, M. Papini, N. Pornputtapong, G. Scalcinati, L. Fagerberg, M. Uhlén, and J. Nielsen, "A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: A case study in Saccharomyces cerevisiae," *Nucleic Acids Research*, vol. 40, no. 20, pp. 10 084–10 097, Nov. 1, 2012.

[166] W. Zhang, Y. Yu, F. Hertwig, J. Thierry-Mieg, W. Zhang, D. Thierry-Mieg, J. Wang, C. Furlanello, V. Devanarayan, J. Cheng, Y. Deng, B. Hero, H. Hong, M. Jia, L. Li, S. M. Lin, Y. Nikolsky, A. Oberthuer, T. Qing, Z. Su, R. Volland, C. Wang, M. D.

Wang, J. Ai, D. Albanese, S. Asgharzadeh, S. Avigad, W. Bao, M. Bessarabova, M. H. Brilliant, B. Brors, M. Chierici, T.-M. Chu, J. Zhang, R. G. Grundy, M. M. He, S. Hebbring, H. L. Kaufman, S. Lababidi, L. J. Lancashire, Y. Li, X. X. Lu, H. Luo, X. Ma, B. Ning, R. Noguera, M. Peifer, J. H. Phan, F. Roels, C. Rosswog, S. Shao, J. Shen, J. Theissen, G. P. Tonini, J. Vandesompele, P.-Y. Wu, W. Xiao, J. Xu, W. Xu, J. Xuan, Y. Yang, Z. Ye, Z. Dong, K. K. Zhang, Y. Yin, C. Zhao, Y. Zheng, R. D. Wolfinger, T. Shi, L. H. Malkas, F. Berthold, J. Wang, W. Tong, L. Shi, Z. Peng, and M. Fischer, "Comparison of RNA-seq and microarray-based models for clinical endpoint prediction," *Genome Biology*, vol. 16, no. 1, p. 133, Jun. 25, 2015.

[167] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[168] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt, "Toward a Shared Vision for Cancer Genomic Data," *New England Journal of Medicine*, vol. 375, no. 12, pp. 1109–1112, Sep. 22, 2016. pmid: 27653561.

[169] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[170] M. Zhao, Y. Tang, H. Kim, and K. Hasegawa, "Machine Learning With K-Means Dimensional Reduction for Predicting Survival Outcomes in Patients With Breast Cancer," *Cancer Informatics*, vol. 17, p. 1 176 935 118 810 215, Jan. 1, 2018.

[171] S. Goli, H. Mahjub, J. Faradmal, H. Mashayekhi, and A.-R. Soltanian. (Nov. 1, 2016). Survival Prediction and Feature Selection in Patients with Breast Cancer Using Support Vector Regression, Computational and Mathematical Methods in Medicine, (visited on 11/18/2020).

[172] D. Sun, A. Li, B. Tang, and M. Wang, "Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome," *Computer Methods and Programs in Biomedicine*, vol. 161, pp. 45–53, Jul. 1, 2018.

[173] N. E. Breslow, "Analysis of survival data under the proportional hazards model," *International Statistical Review/Revue Internationale de Statistique*, pp. 45–57, 1975.

[174] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC medical research methodology*, vol. 18, no. 1, p. 24, 2018.

[175] H. Kvamme, Borgan, and I. Scheel, "Time-to-event prediction with neural networks and cox regression," *Journal of Machine Learning Research*, vol. 20, no. 129, pp. 1–30, 2019.

[176] L. Tong, J. Mitchel, K. Chatlin, and M. D. Wang, "Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 225, Sep. 15, 2020.

[177] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[178] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. Wei, "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in medicine*, vol. 30, no. 10, pp. 1105–1117, 2011.

[179] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, p. 6, 2011.

[180] R.-H. Chung and C.-Y. Kang, "A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification," *GigaScience*, vol. 8, no. 5, May 1, 2019.

[181] C. Martínez-Mira, A. Conesa, and S. Tarazona, "MOSim: Multi-Omics Simulation in R," *bioRxiv*, p. 421 834, Sep. 20, 2018.

[182] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, ser. STOC '98, Dallas, Texas, USA: Association for Computing Machinery, May 23, 1998, pp. 604–613, ISBN: 978-0-89791-962-3.

[183] L. Tong, H. Wu, and M. D. Wang, "Integrating multi-omics data by learning modality invariant representations for improved prediction of overall survival of cancer," *Methods*, Aug. 5, 2020.

[184] M. Goldman, B. Craft, M. Hastie, K. Repečka, F. McDade, A. Kamath, A. Banerjee, Y. Luo, D. Rogers, A. N. Brooks, J. Zhu, and D. Haussler, "The UCSC Xena platform for public and private cancer genomics data visualization and interpretation," *bioRxiv*, p. 326 470, Sep. 26, 2019.

[185] C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami, and A. C. Culhane, "Dimension reduction techniques for the integrative analysis of multi-omics data," *Briefings in Bioinformatics*, vol. 17, no. 4, pp. 628–641, Jul. 1, 2016.

[186] C. Davidson-Pilon, J. Kalderstam, P. Zivich, B. Kuhn, M. Williamson, AbdealiJK, A. Fiore-Gartland, L. Moneda, Gabriel, D. WIlson, A. Parij, K. Stark, S. Anton, M. S. Peña, L. Besson, Jona, H. Gadgil, D. Golland, S. Hussey, R. Kumar, J. Noorbakhsh, A. Klintberg, D. Medvinsky, D. Zgonjanin, D. S. Katz, D. Chen, C. Ahern, C. Fournier, A. Moncada-Torres, and A. F. Rendeiro, *CamDavidsonPilon/lifelines: V0.23.7*, Zenodo, Jan. 14, 2020.

[187] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen, "Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans," *Scientific Reports*, vol. 6, no. 1, p. 24 454, 1 Apr. 15, 2016.

[188] E. E. Bron, M. Smits, W. M. van der Flier, H. Vrenken, F. Barkhof, P. Scheltens, J. M. Papma, R. M. E. Steketee, C. Méndez Orellana, R. Meijboom, M. Pinto, J. R. Meireles, C. Garrett, A. J. Bastos-Leite, A. Abdulkadir, O. Ronneberger, N. Amoroso, R. Bellotti, D. Cárdenas-Peña, A. M. Álvarez Meza, C. V. Dolph, K. M. Iftekharuddin, S. F. Eskildsen, P. Coupé, V. S. Fonov, K. Franke, C. Gaser, C. Ledig, R. Guerrero, T. Tong, K. R. Gray, E. Moradi, J. Tohka, A. Routier, S. Durrleman, A. Sarica, G. Di Fatta, F. Sensi, A. Chincarini, G. M. Smith, Z. V. Stoyanov, L. Sørensen, M. Nielsen, S. Tangaro, P. Inglese, C. Wachinger, M. Reuter, J. C. van Swieten, W. J. Niessen, and S. Klein, "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge," *NeuroImage*, vol. 111, pp. 562–579, May 1, 2015.

[189] L. A. D. Cooper, J. Kong, D. A. Gutman, F. Wang, J. Gao, C. Appin, S. Cholleti, T. Pan, A. Sharma, L. Scarpace, T. Mikkelsen, T. Kurc, C. S. Moreno, D. J. Brat, and J. H. Saltz, "Integrated morphologic analysis for the identification and characterization of disease subtypes," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 317–323, Mar. 1, 2012.

[190] C. S. Carignan and Y. Yagi, "Optical endomicroscopy and the road to real-time, in vivo pathology: Present and future," *Diagnostic Pathology*, vol. 7, no. 1, p. 98, Aug. 13, 2012.

[191] S. J. Spechler and R. F. Souza, "Barrett's esophagus," *The New England Journal of Medicine*, vol. 371, no. 9, pp. 836–845, Aug. 28, 2014. pmid: 25162890.

[192] K. B. Dunbar, P. Okolo, E. Montgomery, and M. I. Canto, "Confocal laser endomicroscopy in Barrett's esophagus and endoscopically inapparent Barrett's neoplasia: A prospective, randomized, double-blind, controlled, crossover trial," *Gastrointestinal Endoscopy*, vol. 70, no. 4, pp. 645–654, Oct. 1, 2009.

[193] M. I. Canto, S. Anandasabapathy, W. Brugge, G. W. Falk, K. B. Dunbar, Z. Zhang, K. Woods, J. A. Almario, U. Schell, J. Goldblum, A. Maitra, E. Montgomery, and R. Kiesslich, "In vivo endomicroscopy improves detection of Barrett's esophagus–related neoplasia: A multicenter international randomized controlled trial (with video)," *Gastrointestinal Endoscopy*, vol. 79, no. 2, pp. 211–221, Feb. 1, 2014.

[194] P. Sharma, A. R. Meining, E. Coron, C. J. Lightdale, H. C. Wolfsen, A. Bansal, M. Bajbouj, J.-P. Galmiche, J. A. Abrams, A. Rastogi, N. Gupta, J. E. Michalek, G. Y. Lauwers, and M. B. Wallace, "Real-time increased detection of neoplastic tissue in Barrett's esophagus with probe-based confocal laser endomicroscopy: Final results of an international multicenter, prospective, randomized, controlled trial," *Gastrointestinal Endoscopy*, vol. 74, no. 3, pp. 465–472, Sep. 1, 2011.

[195] H. Wu, L. Tong, and M. D. Wang, "Improving multi-class classification for endomicroscopic images by semi-supervised learning," in *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, Feb. 2017, pp. 5–8.

[196] L. Tong, H. Wu, and M. D. Wang, "CAESNet: Convolutional AutoEncoder based Semi-supervised Network for improving multiclass classification of endomicroscopic images," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1286–1296, Nov. 1, 2019.

[197] M. B. Sturm and T. D. Wang, "Emerging optical methods for surveillance of Barrett's oesophagus," *Gut*, vol. 64, no. 11, pp. 1816–1823, Nov. 1, 2015. pmid: 25975605.

[198] C. L. Leggett, E. C. Gorospe, D. K. Chan, P. Muppa, V. Owens, T. C. Smyrk, M. Anderson, L. S. Lutzke, G. Tearney, and K. K. Wang, "Comparative diagnostic performance of volumetric laser endomicroscopy and confocal laser endomicroscopy in the detection of dysplasia associated with Barrett's esophagus," *Gastrointestinal Endoscopy*, vol. 83, no. 5, 880–888.e2, May 1, 2016.

[199] D. Kang, S. C. Schlachter, R. W. Carruth, M. Kim, T. Wu, N. Tabatabaei, P. Vacas-Jacques, M. Shishkov, K. Woods, J. S. Sauk, J. Leung, N. S. Nishioka, and G. J. Tearney, "Comprehensive confocal endomicroscopy of the esophagus in vivo," *Endoscopy International Open*, vol. 2, no. 3, E135–E140, Sep. 2014. pmid: 26134959.

[200] C.-Q. Li, X.-L. Zuo, J. Guo, J. Yuan, J.-W. Liu, and Y.-Q. LI, "Sa1492 A Paralleled Comparison Between Two Sets of Confocal LASER Endomicroscopy in Gastrointestinal Tract," *Gastrointestinal Endoscopy*, vol. 79, AB233, May 1, 2014.

[201] P. Sharma, "Clinical practice. Barrett's esophagus," *The New England Journal of Medicine*, vol. 361, no. 26, pp. 2548–2556, Dec. 24, 2009. pmid: 20032324.

[202] S. S. Devesa, W. J. Blot, and J. F. Fraumeni, "Changing patterns in the incidence of esophageal and gastric carcinoma in the United States," *Cancer*, vol. 83, no. 10, pp. 2049–2053, 1998.

[203] J. T. Chang and D. A. Katzka, "Gastroesophageal Reflux Disease, Barrett Esophagus, and Esophageal Adenocarcinoma," *Archives of Internal Medicine*, vol. 164, no. 14, p. 1482, Jul. 26, 2004.

[204] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun, "Cancer Statistics, 2009," *CA: A Cancer Journal for Clinicians*, vol. 59, no. 4, pp. 225–249, 2009.

[205] P. Sharma, K. McQuaid, J. Dent, M. B. Fennerty, R. Sampliner, S. Spechler, A. Cameron, D. Corley, G. Falk, J. Goldblum, J. Hunter, J. Jankowski, L. Lundell, B. Reid, N. J. Shaheen, A. Sonnenberg, K. Wang, and W. Weinstein, "A critical review of the diagnosis and management of Barrett's esophagus: The AGA Chicago Workshop1 1Members of the workshop composed a group of international experts in BE from gastroenterology, surgery, pathology, molecular biology, outcomes, and epidemiology. Conference chairman: Prateek Sharma; conference moderator: Kenneth McQuaid; group leaders: John Dent, M. Brian Fennerty, Richard Sampliner, Stuart Spechler; participants: Alan Cameron, Douglas Corley, Gary Falk, John Goldblum, John Hunter, Janusz Jankowski, Lars Lundell, Brian Reid, Nicholas Shaheen, Amnon Sonnenberg, Kenneth Wang, and Wilfred Weinstein.," *Gastroenterology*, vol. 127, no. 1, pp. 310–330, Jul. 1, 2004.

[206] R. Anaparthy and P. Sharma, "Progression of Barrett oesophagus: Role of endoscopic and histological predictors," *Nature Reviews Gastroenterology & Hepatology*, vol. 11, no. 9, pp. 525–534, 9 Sep. 2014.

[207] K. WANG, "Practice Parameters Committee of the American College of Gastroenterology. Updated guidelines 2008 for the diagnosis, surveillance and therapy of Barrett's esophagus," *Am J Gastroenterol*, vol. 103, pp. 788–797, 2008.

[208] R. S. Gill and R. Singh, "Endoscopic imaging in Barrett's esophagus: Current practice and future applications," *Annals of Gastroenterology*, vol. 25, no. 2, pp. 89–95, 2012. pmid: 24714225.

[209] E. Grisan, E. Veronese, G. Diamantis, C. Trovato, C. Crosta, and G. Battaglia, "239 Computer Aided Diagnosis of Barrett's Esophagus Using Confocal Laser Endomicroscopy: Preliminary Data," *Gastrointestinal Endoscopy*, vol. 75, no. 4, AB126, Apr. 1, 2012.

[210] E. Veronese, E. Grisan, G. Diamantis, G. Battaglia, C. Crosta, and C. Trovato, "Hybrid patch-based and image-wide classification of confocal laser endomicroscopy images in Barrett's esophagus surveillance," in *2013 IEEE 10th International Symposium on Biomedical Imaging*, Apr. 2013, pp. 362–365.

[211] N. Ghatwary, A. Ahmed, X. Ye, and H. Jalab, "Automatic grade classification of Barretts Esophagus through feature enhancement," in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, International Society for Optics and Photonics, Mar. 3, 2017, p. 1 013 433.

[212] R. Mendel, A. Ebigbo, A. Probst, H. Messmann, and C. Palm, "Barrett's Esophagus Analysis Using Convolutional Neural Networks," in *Bildverarbeitung für die Medizin 2017*, g. F. Maier-Hein Klaus Hermann, g. L. Deserno Thomas Martin, H. Handels, and T. Tolxdorff, Eds., ser. Informatik aktuell, Berlin, Heidelberg: Springer, 2017, pp. 80–85, ISBN: 978-3-662-54345-0.

[213] J. Hong, B. Park, and H. Park, "Convolutional neural network classifier for distinguishing Barrett's esophagus and neoplasia endomicroscopy images," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2017, pp. 2892–2895.

[214] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly- and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1742–1750.

[215] Z. Jia, X. Huang, E. I. Chang, and Y. Xu, "Constrained Deep Weak Supervision for Histopathology Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 36, no. 11, pp. 2376–2388, Nov. 2017.

[216] J. Li, W. Speier, K. C. Ho, K. V. Sarma, A. Gertych, B. S. Knudsen, and C. W. Arnold, "An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies," *Computerized Medical Imaging and Graphics*, vol. 69, pp. 125–133, Nov. 1, 2018.

[217] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction," in *Artificial Neural Networks and Machine Learning – ICANN 2011*, T. Honkela, W. Duch, M. Girolami, and S. Kaski, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2011, pp. 52–59, ISBN: 978-3-642-21735-7.

[218] M. D. Bloice, C. Stocker, and A. Holzinger. (Aug. 11, 2017). Augmentor: An Image Augmentation Library for Machine Learning. arXiv: 1708.04680 [cs, stat], (visited on 11/05/2020).

[219] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. (May 24, 2016). Adversarial Autoencoders. arXiv: 1511.05644 [cs], (visited on 11/05/2020).

[220] A. J. Ratner, H. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré, "Learning to Compose Domain-Specific Transformations for Data Augmentation," *Advances in Neural Information Processing Systems*, vol. 30, pp. 3236–3246, 2017.

[221] F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman, "Digital Imaging in Pathology: Whole-Slide Imaging and Beyond," *Annual Review of Pathology: Mechanisms of Disease*, vol. 8, no. 1, pp. 331–359, 2013. pmid: 23157334.

[222] L. Pantanowitz, J. H. Sinard, W. H. Henricks, L. A. Fatheree, A. B. Carter, L. Contis, B. A. Beckwith, A. J. Evans, A. Lal, and A. V. Parwani, "Validating Whole Slide Imaging for Diagnostic Purposes in Pathology: Guideline from the College of American Pathologists Pathology and Laboratory Quality Center," *Archives of Pathology & Laboratory Medicine*, vol. 137, no. 12, pp. 1710–1722, Dec. 1, 2013.

[223] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255.

[224] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2014, pp. 740–755, ISBN: 978-3-319-10602-1.

[225] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[226] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.

[227] Z. Xu and Q. Zhang, "Multi-scale context-aware networks for quantitative assessment of colorectal liver metastases," in *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, Mar. 2018, pp. 369–372.

[228] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[229] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, M. Kohl, C. Walz, F. Ludwig, S. Braunewell, M. Baust, Q. D. Vu, M. N. N. To, E. Kim, J. T. Kwak, S. Galal, V. Sanchez-Freire, N. Brancati, M. Frucci, D. Riccio, Y. Wang, L. Sun,

K. Ma, J. Fang, I. Kone, L. Boulmane, A. Campilho, C. Eloy, A. Polónia, and P. Aguiar, "BACH: Grand challenge on breast cancer histology images," *Medical Image Analysis*, vol. 56, pp. 122–139, Aug. 1, 2019.

[230] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 8, pp. 1962–1971, Aug. 2016.

[231] L. Tong, Y. Sha, and M. D. Wang, "Improving Classification of Breast Cancer by Utilizing the Image Pyramids of Whole-Slide Imaging and Multi-scale Convolutional Neural Networks," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 1, Jul. 2019, pp. 696–703.